

Syllabus

PANJAB UNIVERSITY, CHANDIGARH

INTRODUCTION TO ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

BCA 2nd Semester (NEP)

Paper Code: BCA-202

L	T	P	Cr
2	-	4	4

Theory External Marks: 45
Theory Internal Marks: 05
Number of Theory Hours : 30
Practical Ext. Marks : 50

OBJECTIVE :

The course will help students to learn the basics of Artificial Intelligence and Data Science. It further aims to make them analyze data using spreadsheets.

COURSE LEARNING OUTCOMES

On successful completion of this course, a student will be able to

(*i*) understand basics of Artificial Intelligence and Data Science.

(*ii*) analyze data using spreadsheet software.

(*iii*) visualize, search and summarize the data using spreadsheet.

Note: (*i*) The Question Paper will consist of Four Units.

(*ii*) Examiner will set total of NINE questions comprising TWO questions from each Unit and ONE compulsory question of short answer type covering whole syllabi.

(*iii*) The students are required to attempt ONE question from each Unit and the Compulsory question.

(*iv*) All questions carry equal marks unless specified.

UNIT-I

Introduction to Artificial Intelligence (AI) : Definition, Types of AI, History of AI, Applications of AI in healthcare, education and agriculture, Introduction to key technologies of AI: Machine Learning, Deep Learning, Natural Language Processing (NLP), Computer Vision, Robotics; Ethical issues and Future of AI. Problem Solving: AI problems, Agents and Environments, Structure of Agents.

UNIT-II

Data Science: Introduction to Data Science, evolution and need for Data Science, Types of Data, Data Science Applications in Various Fields, Data Science Life Cycle or Project Stages, Data collection, Data Pre-Processing Overview – Data Cleaning, Data Integration and Transformation, Data Reduction – Data Discretization, Data Storage and management, careers in Data Science

UNIT-III

Managing Datawith Spreadsheet Software: Creating a database, sorting& filtering data, custom sorting and advanced filtering, definingand grouping database ranges, applying data tools like data validation, removing duplicates, text to columns, Flash fill; Managing data using Math, statistical, Trigonometric, String/Text, Date and Time, logical and Database functions, Nested functions.

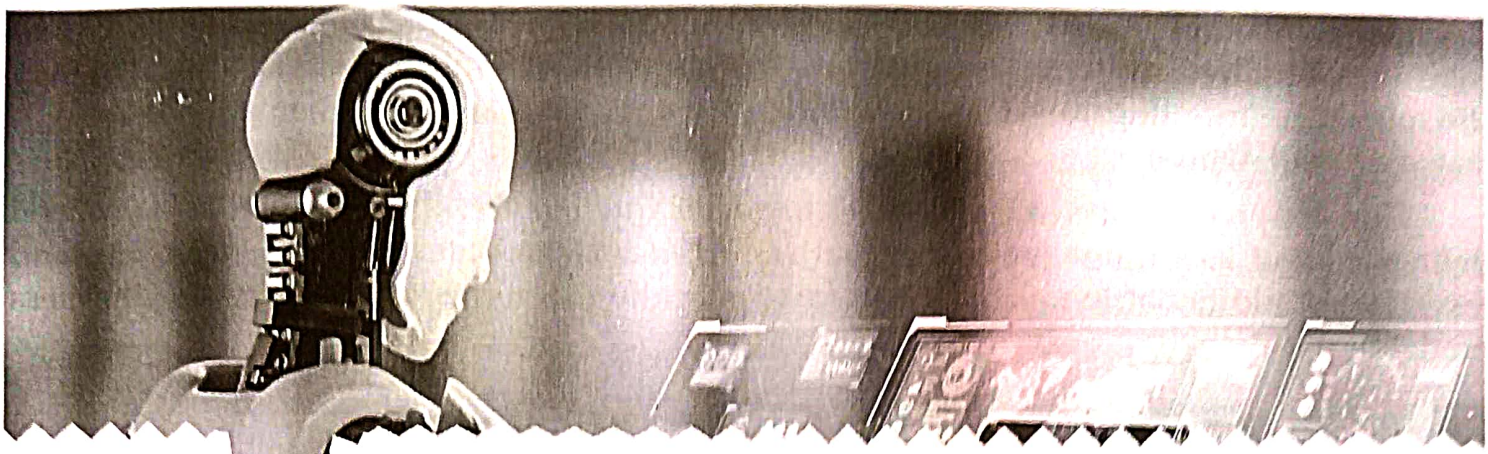
UNIT-IV

Data Visualization, Searching and Summarizing using Spreadsheet Software: Understanding Chart types, Creating Embedded Charts, formatting Charts: Changing Chart types, adding Titles, Legends and Gridlines, Printing Charts, searching data using HLOOKUP and VLOOKUP, summarizing data using pivot table, analyzing data using What-if Analysis: goal seek, scenario manager and data Table.

Contents



Ch.No.	Chapter	Pages
1.	INTRODUCTION TO ARTIFICIAL INTELLIGENCE	1.1–1.26
2.	INTRODUCTION TO KEY TECHNOLOGIES OF AI	2.1–2.33
3.	PROBLEM SOLVING AND APPLICATIONS OF AI	3.1–3.35
4.	INTRODUCTION TO DATA SCIENCE	4.1–4.23
5.	DATA PRE-PROCESSING	5.1–5.30
6.	SORTING AND FILTERING DATA	6.1–6.32
7.	DATA RANGE AND DATA TOOLS	7.1–7.27
8.	MATH, STATISTIC AND TRIGONOMETRICEXCEL FUNCTION	8.1–8.46
9.	STRING, DATE & TIME AND LOGICAL EXCEL FUNCTION	9.1–9.28
10.	CHARTS AND PIVOT TABLE	10.1–10.57
11.	SEARCHING USING HLOOK & VLOOK AND ANALYZING DATA	11.1–11.26



CHAPTER

1 INTRODUCTION TO ARTIFICIAL INTELLIGENCE

CHAPTER AT A GLANCE

1.1	Introduction	1.6	Advantages of AI
1.2	What is Artificial Intelligence??	1.7	Disadvantages of AI
1.3	Goals of Artificial Intelligence	1.8	Components of Artificial Intelligence
1.4	Component Disciplines of AI	1.9	Types of AI
1.5	Features of AI	1.10	History of AI

1.1 INTRODUCTION

People have been developing tools since time immemorial. During prehistoric times, we used stones to make harpoons for hunting. During the early days of agriculture, we developed tools such as plows and sickles, carts for big loads, and even equipment for making additional tools. Over time, they become increasingly intricate. Then we invented machines. Machines are more than just tools; they use energy to complete things that people would struggle with. The vehicles, washing machines, mixers, lawn mower, etc. have now become the indispensable inventions.

Humans are skilled at problem solving and making judgments. Humans are capable of making their own decisions. This explains why humans are considered intelligent. Human intelligence can be adapted and applied to machines. We can create computers and devices with human-like intelligence. This is referred to as artificial intelligence. Machines, once limited to performing repetitive tasks, are now equipped with intelligence (Artificial Intelligence) that allows them to learn, reason, and make decisions. This transformation is reshaping industries, communication, and daily life.Artificial intelligence is sometimes described as a new revolution that is a fast-growing branch of computer science, a discipline famous for its interdisciplinarity and free-spirited interest in next-generation challenges of automation. Artificial intelligence encompasses a variety of

disciplines, making it an expansive field. Developing intelligent computer systems and machines requires drawing insights and concepts from multiple areas of study. Key disciplines contributing to AI include biology, computer science, mathematics, sociology, statistics, and more.

This chapter discuss the basics of artificial intelligence. The chapter explains and defines AI and its various types, summarizes AI history and literature advancements, and investigates additional innovation that AI is providing in terms of scientific and business models.

1.2 WHAT IS ARTIFICIAL INTELLIGENCE??

Artificial Intelligence is the subfield of computer science that studies the creation of machines or computers with intellect comparable to that of humans. Studying artificial intelligence includes how the human brain functions, how people learn, how they make choices, and how they solve issues. The results of this kind of research are subsequently applied to the creation of intelligent systems and software.When a machine demonstrates human-like traits like reasoning, learning, and problem-solving, we say we have artificial intelligence. In artificial intelligence, a machine can be built using preprogrammed algorithms that operate on its own intelligence instead of needing to be preprogrammed to perform a certain task.

Artificial Intelligence (AI) is focused on creating machines capable of mimicking cognitive functions typically associated with human intelligence. These include learning, reasoning, problem-solving, perception, and language understanding. The two parts of the term artificial intelligence (AI) can be broken down to understand it:

- **Artificial:** Anything that is manufactured or engineered by humans as opposed to existing naturally is referred to as artificial. It indicates systems or items built with a specific purpose in mind.
- **Intelligence:** Intelligence is the ability to acquire and apply knowledge, solve problems, reason, and adapt to new situations. Cognitive processes including learning, comprehension, and decision-making are linked to it.

Therefore, systems created by humans that are intended to replicate the cognitive capacities of natural intelligence are referred to as artificial intelligence.

Some of the definitions of AI given by famous authors are given as:

- **John McCarthy, a pioneer in AI, who coined the term "Artificial Intelligence" in 1956, defines it as:** *"The science and engineering of making intelligent machines, especially intelligent computer programs."*

  McCarthy emphasized the technological and scientific approach to replicating human-like thought processes.

- **Marvin Minsky (1968) :** *"Artificial intelligence is the science of making machines do things that would require intelligence if done by men."*

This definition highlights the goal of enabling machines to perform tasks traditionally requiring human intelligence.

- **Elaine Rich, in her book Artificial Intelligence (1983) , defines artificial intelligence as:** *"The study of how to make computers do things which, at the moment, people do better."*

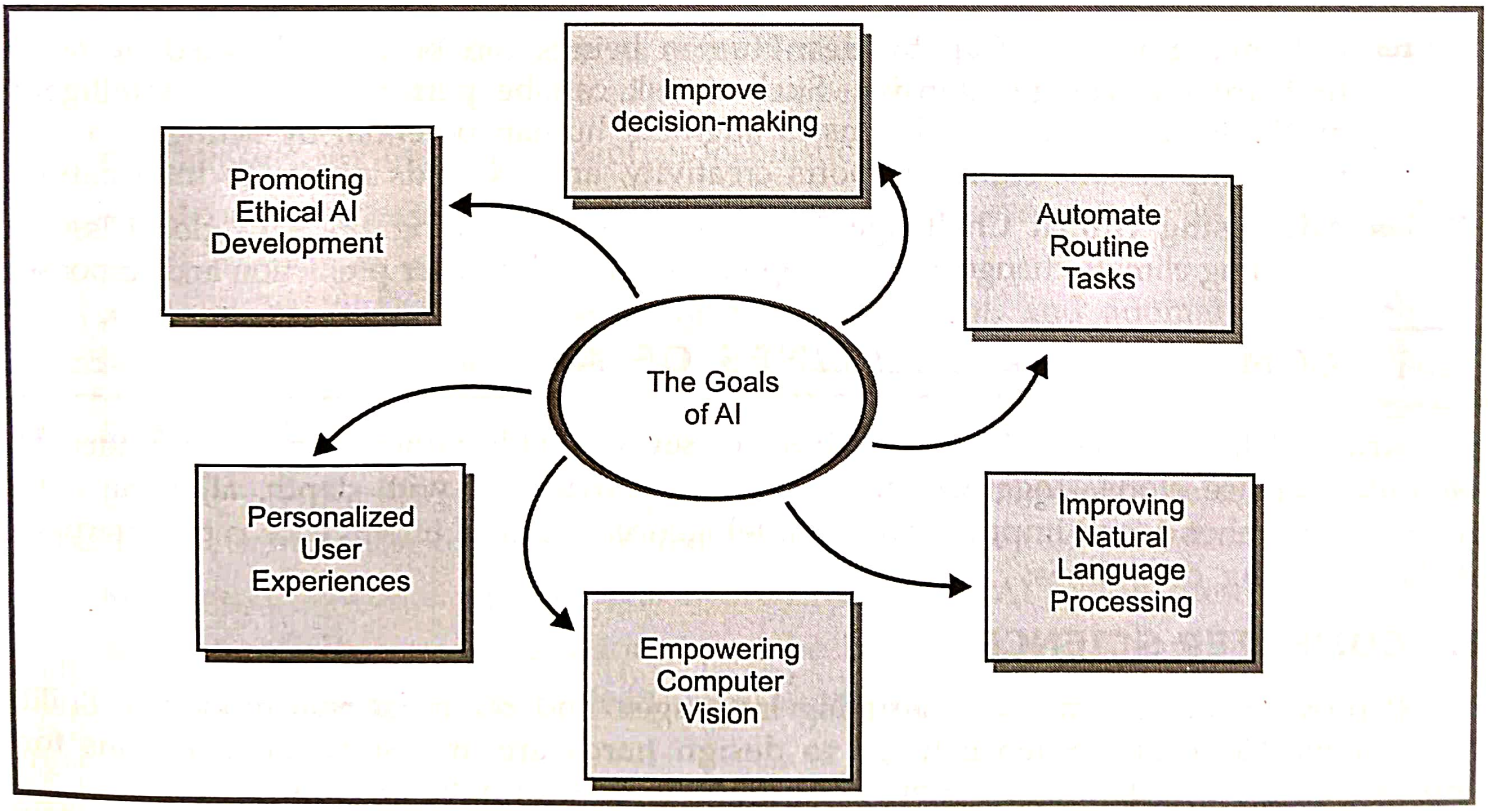
This definition highlights AI's goal of bridging the gap between human intelligence and machine capabilities, pushing the boundaries of what computers can achieve.

- **Stuart Russell and Peter Norvig (1995) , authors of 'Artificial Intelligence: A Modern Approach', describe AI as:** *"The study of agents that perceive the world around them and take actions that maximize their chances of success."*

This modern view focuses on AI as intelligent agents capable of adaptive behavior.

1.3 GOALS OF ARTIFICIAL INTELLIGENCE

Since its conception, the goal of artificial intelligence research has been to produce robots capable of doing activities that were previously limited to human capabilities, such as decision-making, natural language processing, and complicated problem-solving. Fast forward to now, and the field of AI is characterized by an array of applications, ranging from self-driving cars to virtual help, ushering in significant improvements in a variety of areas such as medicine and healthcare, finance, and manufacturing. Today, artificial intelligence shows symptoms of not only automating performance but also having the potential to revolutionize the way we live and work.



Let us explore the various goals of artificial intelligence in greater detail below:

- **Automation of Repetitive Tasks:** AI seeks to automate mundane, repetitive, or time-consuming tasks, improving efficiency and reducing human effort. Examples include data entry, email filtering, and assembly line operations.

- **Human-Like Intelligence:** Researchers want to create systems that understand, think, learn, and behave like humans. A significant goal is to develop systems that mimic human thinking, learning, and decision-making processes, enabling machines to solve problems creatively and adapt to new situations.
- **Natural Interaction:** AI strives to enable machines to interact naturally with humans, using speech, gestures, or written language. This goal is evident in virtual assistants like Siri or Alexa and natural language processing systems like chatbots.
- **Data Analysis and Insights: By** analyzing large datasets, AI aims to identify patterns, make predictions, and generate actionable insights, aiding fields like business, healthcare, and science.
- **Personalization:** AI works to tailor user experiences based on individual preferences and behavior, enhancing applications like recommendation systems and adaptive learning platforms.
- **Problem-Solving and Decision-Making:** AI aspires to assist in solving complex problems and making decisions, whether through expert systems in medical diagnostics or AI-driven decision models in finance. Artificial intelligent-powered machines can show intelligent behavior such as learning new things on their own and advising the user.
- **Autonomy in Systems:** A crucial goal is to develop autonomous systems, such as self-driving cars or drones, capable of operating independently in dynamic environments.
- **Enhancing Human Capabilities:** Human beings can be overwhelmed by tasks that are knowledge-intensive. Such a task can be performed using intelligent machines and systems. AI aims to augment human potential by acting as a tool that complements skills, supports creativity, and expands access to information.
- **Addressing Global Challenges:** AI focuses on addressing pressing global issues, including climate change, resource optimization, and disaster prediction and response.

1.4 COMPONENT DISCIPLINES OF AI

Artificial Intelligence (AI) is a synthesis of several fields rather than being limited to computer science alone. Together, these domains provide AI with depth, allowing it to imitate, comprehend, and improve human intelligence. Artificial intelligence is the interplay of the following disciplines:

1. COMPUTER SCIENCE

It provides algorithms, programming languages, and computational models to build AI systems. Computer Science helps to design hardware and software platforms for implementing AI technologies. Key Contributions of Computer Science include:

- Development of machine learning frameworks and programming languages (e.g., Python, TensorFlow).
- Design of algorithms for data processing, optimization, and automation.
- Implementation of neural networks and deep learning architectures.

Computer Science is being successfully applied in the field of image recognition, robotics, and autonomous systems that rely on efficient programming and software engineering.

2. MATHEMATICS

Mathematics supports AI by providing the theoretical and practical tools required for data analysis, model construction, and decision-making. Key contributions of Mathematics include:

- Linear algebra enables neural networks and transformations.
- Probability and statistics are used to make predictions, identify patterns, and model uncertainty.
- Calculus enables optimization techniques such as gradient descent in machine learning.
- Graph theory is useful for representing knowledge and analyzing networks.

Mathematical models are essential for reinforcement learning, natural language processing, and recommendation systems.

3. STATISTICS

Statistics is a critical component of AI, allowing for data-driven decision-making, predictive modelling, and hypothesis testing. Key contributions include:

- Statistical approaches such as regression, clustering, and classification, which are essential for both supervised and unsupervised learning.
- Bayesian networks represent probabilistic relationships between variables.
- Hypothesis testing and statistical inference help to assess and ensure the reliability of AI models.

AI applications such as forecasting, risk assessment, and anomaly detection rely significantly on statistical approaches.

4. BIOLOGY

Biology, particularly neuroscience, has had a substantial impact on AI by generating models that imitate the operation of the human brain. Key contributions include:

- Neural networks replicate real neurons, allowing AI to learn and adapt.
- Natural selection processes serve as the basis for evolutionary algorithms.
- Swarm intelligence is derived from biological systems such as ant colonies and bird flocking.

Healthcare AI applies biological concepts to diagnostics, drug discovery, and genomics.

5. PSYCHOLOGY

Psychology gives insights into human cognition, learning, and behavior, which serve as the foundation for developing AI systems that replicate human mental processes.

Key contributions include

- Understanding how humans absorb information, which influences cognitive architectures in AI.
- Reinforcement learning, which uses rewards and penalties to teach computers, is influenced by behavioral psychology.
- Human-computer interface (HCI) research improves AI usability.

AI chatbots and virtual assistants, such as Siri, mimic genuine conversations through the usage of psychological concepts.

6. NEURONS STUDY

Neuroscience has a direct impact on the development of AI, notably in the creation of systems that replicate human cognition and brain activity. Key contributions include:

- **Artificial Neural Network (ANN) :** Machine learning is built on the foundation of ANNs, which are modelled on biological neurons.
- Neuroscience concepts like as memory, attention, and perception influence AI's capacity to process data and prioritize tasks.
- Brain-machine interfaces seek to integrate AI directly into human neural systems.

Neuron Study finds its applications in the field of AI in Neuromorphic computing, which creates hardware modeled after the brain, enables quicker and more energy-efficient AI systems.

7. SOCIOLOGY

Sociology investigates the social consequences and human-centered design of artificial intelligence systems, assuring their ethical and equitable incorporation into society. Key contributions include:

- Understanding human behavior and group dynamics enhances AI design for collaborative systems.
- Ethical AI research aims to reduce bias, ensure fairness, and promote inclusivity.
- Social network analysis helps to analyze online behavior and information diffusion.

AI in social media platforms utilizes sociological principles to predict user behavior and increase engagement.

8. ENGINEERING

Engineering plays a crucial role in applying AI technologies to real-world systems.

Key contributions include:

- Development of AI computation hardware (e.g., GPUs).
- AI integration with robots, autonomous systems, and Internet of Things devices.
- Ensuring scalability and dependability of AI solutions.

The development of AI is a collaborative endeavor that combines knowledge and techniques from numerous areas. This multidisciplinary approach assures that AI systems

are not just clever, but also ethical, efficient, and effective in solving real-world problems. AI continues to push the limits of what machines can do by combining discoveries from computer science, mathematics, psychology, biology, and other fields. AI comes at the junction of these fields, with their combined strengths enhancing its ability to learn, reason, and adapt. For example:

- Machine learning is based on mathematics and statistics, while its implementation relies on computer science.
- Natural language processing (NLP) mixes computer science, psychology, and sociology to construct systems that comprehend and generate human discourse.
- Robotics combines biology, neuroscience, and computer science to help machines perceive and interact with their surroundings.

1.5 FEATURES OF AI

Artificial intelligence has evolved beyond the well-known ideas of science fiction to become a transformational force that improves people's lives, transforms industries, and pushes the limits of technology's potential. Initially based on complex algorithms and computer capacity, AI expanded into multidimensionality in its potential by performing activities that appeared to require human intelligence. AI is transforming the way we interact with technology and how businesses operate, from regular task automation to deep data analytics. This technology's long-term learnability, adaptability, and road to constant progress place it at the center of modern technology. The core features of AI can be given as follows:

1. **Learning:** AI systems can learn from data, improving their performance over time without being explicitly programmed. This feature is primarily realized through machine learning (ML) algorithms.
2. **Reasoning and Problem-Solving:** AI can reason logically and solve complex problems by applying algorithms, heuristics, and knowledge bases.
3. **Natural Language Processing (NLP) :** NLP enables AI to understand, interpret, and generate human language in a way that is both meaningful and contextually appropriate.
4. **Perception:** AI systems can interpret and understand sensory information from their environment (e.g., through visual, auditory, or other sensory inputs).
5. **Autonomy:** AI systems are capable of performing tasks and making decisions without human intervention, often relying on sensors and algorithms.
6. **Adaptability:** AI systems can adjust and evolve their behavior based on new data, experiences, or changes in the environment.
7. **Knowledge Representation:** AI uses structured formats to represent and organize information, allowing systems to reason, understand, and make decisions based on that knowledge.

8. **Decision-Making:** AI can analyze data and make decisions based on predefined rules or learned patterns. This decision-making process can involve complex trade-offs and evaluations.
9. **Collaboration and Interaction:** AI systems can collaborate with humans or other AI systems to enhance productivity and solve problems collectively.
10. **Ethical and Transparent Functioning:** Ethical AI is designed to minimize biases, ensure transparency, and operate responsibly within societal and legal frameworks.

Alternately in this section, we'll look at the essential features of artificial intelligence in plain and straightforward language, allowing beginners to understand why AI is such an influential technology.

1. Automating Repetitive and Monotonous Tasks

Artificial intelligence (AI) systems excel in automating repetitive and tedious processes, reducing the need for human intervention. This not only increases productivity through improved accuracy, but it also allows humans to concentrate on more creative and difficult problem-solving activities. Here's a closer look at how AI does this:

- **Automate repetitive tasks:** AI systems are extremely efficient in managing repetitive and routine tasks, which are frequently time-consuming and prone to human error. These jobs might range from simple data entry to complex operations across a variety of businesses.
- **Data Entry and Management:** Data entry is an excellent example of a tedious work that AI can effectively automate. AI-powered systems process and input data with high precision while minimizing errors. Unlike humans, which can fatigue and make mistakes, AI systems function with a high degree of accuracy that is crucial for maintaining data integrity.

  Example: AI-driven Optical Character Recognition (OCR) technology automates the scanning and digitization of paper documents, converting them into editable and searchable digital formats. This automation not only saves time but also significantly reduces errors in data processing.
- **Increasing Accuracy and Productivity:** AI significantly increases productivity by automating repetitive processes. Employees can now complete activities that truly call for human ingenuity and creativity in minutes instead of hours. This change from routine to strategic work likely leads to increased job satisfaction and fosters an environment that encourages creativity.
- **Automation of robotic processes (RPA) :** One area of artificial intelligence that deals with process automation in businesses is robotic process automation. RPA uses machine learning and artificial intelligence to handle repetitive, high-volume tasks that normally call for human interaction.

2. Human Intelligence Imitation

The ability of machines to mimic human brain activity is among the most amazing aspects of artificial intelligence. Artificial intelligence (AI) systems may be able to learn

from data, recognize patterns resembling those in the human brain, and then make decisions based on prior experiences thanks to sophisticated techniques like machine learning via neural networks. Because of its cognitive ability, AI can handle a variety of difficult jobs with extreme precision, revolutionizing entire industries and, in turn, impacting day-to-day living. Let's take a closer look at how AI is mimicking human thought processes and consider the implications of this capability.

- **Learning from Data:** Machine learning is at the heart of AI's cognitive skills, because operating systems learn from data without explicit programming. Large amounts of data may be processed by AI models, which can then recognize patterns and draw conclusions from the data to make predictions and judgments. For instance, in language processing, artificial intelligence (AI) systems examine vast amounts of text to understand the intricacies of the human language, which allows them to do tasks like summarization and translation with remarkable precision.
- **Pattern Recognition and Decision Making:** AI's cognitive processes are powered by pattern recognition. For instance, by learning from annotated photographs, AI systems in image analysis are able to identify faces, entities, scenes, and other items in images. Similar to this, AI in voice recognition can identify words and phrases in audio recordings. AI is able to make informed decisions because of its pattern recognition. In order to operate safely and effectively, self-driving cars, for example, are outfitted with artificial intelligence (AI) that can identify pedestrians, traffic signals, and other vehicles.
- **Natural language processing, or NLP:** It is one area of artificial intelligence research that focuses on how computers and human language interact. AI systems can speak, respond to inquiries, and suggest entities when they are able to comprehend and produce human language. This is how natural language processing (NLP) is used by virtual assistants such as Siri, Alexa, or Google Assistant to comprehend voice requests and respond appropriately. They display some of AI's cognitive capabilities by being able to play music, create reminders, check the weather, and even operate smart gadgets at home.

3. Data Ingestion

The process by which AI systems gather and analyze enormous volumes of data quickly and accurately from many sources is known as data intake. Large datasets are necessary for AI systems to produce precise forecasts and judgments. This capability enables AI to collect information from various sources, including sensors, databases, social media, and human interactions, and utilize it to enhance its functionality. These days, a plethora of information from diverse sources surrounds every company across many industries.

The complete analysis of the role and applications of data ingestion in AI, can be given as follows:

- **Speed and Efficiency in Data Processing:** AI systems are made to process large amounts of data, which human analysts would find daunting and time-consuming. In essence, data ingestion entails gathering vast amounts of data from multiple

sources, converting them into a format that can be used, and then importing them into a database or data warehouse for analysis.

- **Data Collection:** AI may use data from real-time streams, databases, cloud storage, and APIs, among other sources. As an additional illustration, data in the financial industry may originate from social media trends, financial reports, and feeds on the stock market. EHRs, medical devices, and clinical trials are some of the sources of data in the health sector.
- **Data Transformation:** After data collection, it will need to be cleaned and formatted for analysis. AI algorithms can assist in automating quality control procedures that guarantee the accuracy and consistency of the data, such as addressing missing values, standardizing formats, and checking for duplicates. For example, in marketing, data from multiple channels and campaigns may be standardized to provide a single perspective of all client interactions.
- **Data Loading:** Loading the data into a database or data warehouse is the final stage of the ingestion process. AI can make sure that everything runs well so that retrieval and storage are successful.
- **Analysis and Interpretation:** This will enable AI systems that have been fed data to examine and decipher the data in order to derive insightful conclusions. This is especially necessary in sectors where prompt and accurate data analysis is the tried-and-true method of making decisions.

4. Face Recognition

AI facial recognition uses deep learning algorithms to verify a person's identity based on facial features in images or videos. This technology founded on cutting-edge AI algorithms, aim to automate processes that would otherwise be done manually, improving security, customer experience, and overall user experience. Some common examples can be security systems, smartphones, and social media applications for tasks like unlocking devices, tagging photos, etc.

5. Chatbots

Chatbots are artificial intelligence-powered virtual assistants designed to communicate with humans. They comprehend and react to the user's queries in a conversational manner, utilizing natural language and machine learning features. There are two types of Chatbots:

- **Rule-Based Chatbots:** These work using predefined rules and scripts, making simple inquiries and providing particular responses to keywords or phrases. The entire process of a rule-based chatbot is thus efficient for simple actions that follow its predefined rules.
- **AI-Based Chatbots:** These chatbots employ an NLP and machine learning system to understand context and intent from user questions. These can progress to more complex conversations, be aware of the learning from exchanges, and provide more correct and appropriate responses.

6. Cloud Computing

The ability of AI to work at scale is heavily reliant on cloud computing. Artificial intelligence (AI) applications sometimes require massive quantities of processing power and data storage, which cloud systems such as Microsoft Azure, Amazon Web Services (AWS) , and Google Cloud deliver. These platforms let organizations and developers to access scale resources without investing in costly infrastructure.

- **Benefits of Cloud Computing for AI:** Cloud systems provide the high-performance computing resources needed for applications such as training massive AI models.
- **Storage:** Cloud solutions provide the easy storage and access of massive volumes of data used by AI systems.
- **Accessibility:** Cloud computing makes AI technology more accessible to smaller enterprises, allowing them to harness AI without the requirement for their own data centers.

1.6 ADVANTAGES OF AI

1. **Automation of Tasks:** AI systems can eliminate repetitive, time-consuming interventions and save time by automating manual tasks. Automation improves efficiency and productivity, minimizes human error and frees the humans for more work.
2. **Enhanced decision-making:** AI systems can digest enormous amounts of data, understand patterns, and make informed decisions more quickly than humans. This increases predictive and diagnostic accuracy and facilitates sophisticated decision-making, such as financial forecasting and medical diagnostics.
3. **Increased Efficiency and Productivity:** Artificial intelligence accelerates operations and lowers downtime in industries, increasing total production. AI optimizes workflows and resource use which improves output quality.
4. **Personalization:** AI systems can assess user preferences and habits to deliver personalized recommendations or services. Personalization helps in improved user happiness and consumer experience.
5. **Cost Savings:** Artificial intelligence decreases operating expenses by automating operations and decreasing errors. AI reduces labor costs on repetitive jobs, inefficiency and wastage.
6. **Efficient handling of Complex and Large-Scale Data:** AI systems excel in managing and processing large volumes of data, providing insights that would otherwise be difficult to obtain.
7. **24X7 Availability and Reliability:** In contrast to human labor, AI systems can perform continuously without becoming fatigued. AI systems ensures consistency in service delivery and handles heavy workloads effectively.

8. **Innovation and Creativity:** AI promotes innovation by creating new concepts, designs, and solutions. It promotes creativity in domains such as art, music, and writing.
9. **Enhanced Safety and Fraud Detection:** AI enhances fraud detection and prevention by analyzing transaction patterns and identifying anomalies that may indicate fraudulent activities. Machine learning algorithms can detect unusual behavior and flag suspicious transactions in real time, allowing organizations to take immediate action.
10. **Bias and Fairness:** AI can identify and reduce bias in decision-making, improving fairness and equality. By analyzing massive datasets, AI can identify bias tendencies and provide insights into how they affect outcomes. Furthermore, AI algorithms can be built to reduce biases, ensuring that decisions are made based on objective criteria rather than subjective or biased ones.
11. **Scalability:** AI not only operates continuously, but it can scale virtually endlessly. Amazon and Netflix, for example, provide individualized recommendations to their clients. While a sales clerk who interacts with a customer on a regular basis may be able to provide such services to that individual, AI can do so for hundreds of thousands of customers at once by evaluating existing customer data. AI displays similar scalability in the finance industry, where institutions employ the technology to instantaneously evaluate and authenticate millions of transactions and monitor for potential fraud on a daily basis.
12. **Improved accuracy and lowered error rate :** Unlike humans, AI systems can handle exponentially more information and consistently follow the rules for analyzing data and making judgments, making them far more likely to produce accurate findings almost every time.

    However, there is one major limitation. To achieve such accuracy, AI models must be based on strong algorithms that are free of unintended bias, trained on sufficient high-quality data, and monitored for drift.
13. **Absence of emotion and judgment:** AI lacks human emotions and judgment, making it an effective tool in a number of situations. When dealing with confused customers, AI-enabled customer care chatbots will not grow agitated, pass judgment, or become aggressive.

1.7 DISADVANTAGES OF AI

1. **Over Dependence on Technology:** People who rely too much on AI risk losing their ability to understand complex problems and circumstances. Although this disadvantage only applies when AI is utilized regularly, it is undoubtedly possible.

   In short, individuals should not rely solely on AI and rather make good use of their own intelligence. Even with so many new technical options, you still need personnel that have a thorough understanding of processes and workflows.
2. **Job loss and displacement:** According to a Goldman Sachs Research paper from April 2023, artificial intelligence "could expose the equivalent of 300 million

full-time jobs to automation." The scientists also projected that "roughly two-thirds of U.S. occupations are subject to some degree of automation by AI."

Jobs in manufacturing, retail, customer service, and even certain professional fields such as legal research or medical diagnostics are progressively becoming automated, resulting in major job displacement.

While AI can open up new career prospects, the transition period can be difficult, with many workers needing retraining and upskilling. If significant job displacement is not adequately handled, it has the potential to raise unemployment and social inequality.

3. **Costly implementation:** The most significant and obvious disadvantage of integrating AI is that its development can be exceedingly expensive. The cost of using AI varies depending on what you need it for.
4. **Data security and privacy:** Sharing data with an AI system allows it to learn and grow. It only becomes problematic if the data is excessively personal, is only for internal use, or contains other sensitive information. Users should therefore think about what information they put into an AI system and what they don't. After all, they are sharing the data with the corporation that owns the database, as well as, if necessary, other users.
5. **Lack of Creativity:** AI frequently lacks the inherent creativity of humans, which is derived from emotional depth, abstract reasoning, and inventive processes. While AI can mimic creativity by creating art, music, or writing based on pre-existing patterns, it lacks actual originality and the ability to think outside the box.
6. **Emotional Intelligence:** The next problem of AI is that it lacks emotional intelligence, which requires detecting and managing one's own emotions, as well as empathizing with others and treating interpersonal connections with prudence and empathy.
7. **Ethical problems:** The fast development and application of AI raised numerous ethical concerns about its use and future growth. One of the most prevalent ethical issues raised is worry over consumer data privacy. The durability of data creates numerous challenges for the informed consent of the individuals to whom the data belongs. Furthermore, because AI is strong at spotting patterns, it can collect data on people without direct access to personal information.
8. **High energy requirements and questions of sustainability:** Although AI can and is now being used to assist firms in developing more sustainable practices, some experts have raised concern that AI's energy requirements may harm rather than benefit sustainability initiatives, especially in the short term.
9. **Lacks knowledge of Intricacies of AI Algorithm:** AI algorithms can be comparable to black boxes. Very less is known about the internal working of an AI algorithm. We only know of the prediction the system produces but is not aware of how the system arrives at this prediction. This gives the issue of being trustworthy of the AI-generated tasks or products.

Artificial intelligence (AI) is a strong technology that presents both enormous benefits and considerable concerns. On the positive side, AI improves efficiency, automates tedious

processes, and makes precise data-driven decisions, making it invaluable in industries such as healthcare, banking, and transportation. It promotes innovation, resulting in advances such as tailored medicine and self-driving cars. AI also increases accessibility by providing solutions for people with disabilities via technologies such as speech-to-text and visual aid. However, its implementation raises ethical questions about bias, data privacy, and responsibility, as well as the possibility of job displacement owing to automated processes. The high cost of developing and maintaining AI systems is also a deterrent, especially for smaller firms. Furthermore, the possibility of misuse, such as deepfakes or AI-powered cyberattacks, raises worries about security and misinformation.

1.8 COMPONENTS OF ARTIFICIAL INTELLIGENCE

Any AI system is primarily composed of the following components: a learning AI programming language, knowledge representation, problem solving (mostly through heuristic search) , and AI hardware. Let's take an overview of them:

(a) Learning: Learning means adding new knowledge to the knowledge base and improving or refining previous knowledge. The success of an AI program is based on the extent of knowledge it has and how frequently it acquires knowledge. Learning agents consists of four main components. They are the:

- **Learning element**, the part of the agent responsible for improving its performance.
- **Performance element**, the part that chooses the actions to take.
- **Critics**, which tells the learning element how the agent is doing.
- **Problem generator**, which suggests actions that could lead to new information experiences.

(b) AI Programming Language: Today, we have specialized languages and programs for data processing and scientific applications, as well as specialized languages and tools for AI programming that use AI language programs and AI environment tools. LISP and Prolog are the most commonly used languages in AI development.

- **LISP (List Processing) :** LISP is an AI programming language developed by John McCarthy in 1950. LISP is a symbolic processing language that represents information in lists and manipulates lists to derive information.
- **PROLOG (Programming in Logic) :** Prolog was developed by Alain Colmeraver and P. Roussel at Marseilles University in Francein the early 1970's. Prolog uses the syntax of predicate logic to perform symbolic, logical computations.

(c) Knowledge Representation: The quality of the result depends on how much knowledge the system possesses. The available knowledge must be represented in an efficient way. Hence, knowledge representation is a vital component of the system. The best known representations schemes are:

- Associative Networks or Semantic Networks
- Frames

- Conceptual Dependencies and
- Scripts

(d) Problem-solving: The objective of this particular area of research is how to implement the procedures on AI systems to solve problems like humans do. The inference process should also be equally good to obtain satisfactory results. The inference process is broadly divided into the brute and heuristic search procedures.

(e) AI Hardware: Today, most of the AI programs are implemented on Von Neumann machines only. However, dedicated workstations have emerged for AI programming. Computers are classified into one of following four categories:

(a) Single Instruction Single Data (SISD) Machines

(b) Single Instruction Multiple Data (SIMD) Machines

(c) Multiple Instruction Single Data (MISD) Machines

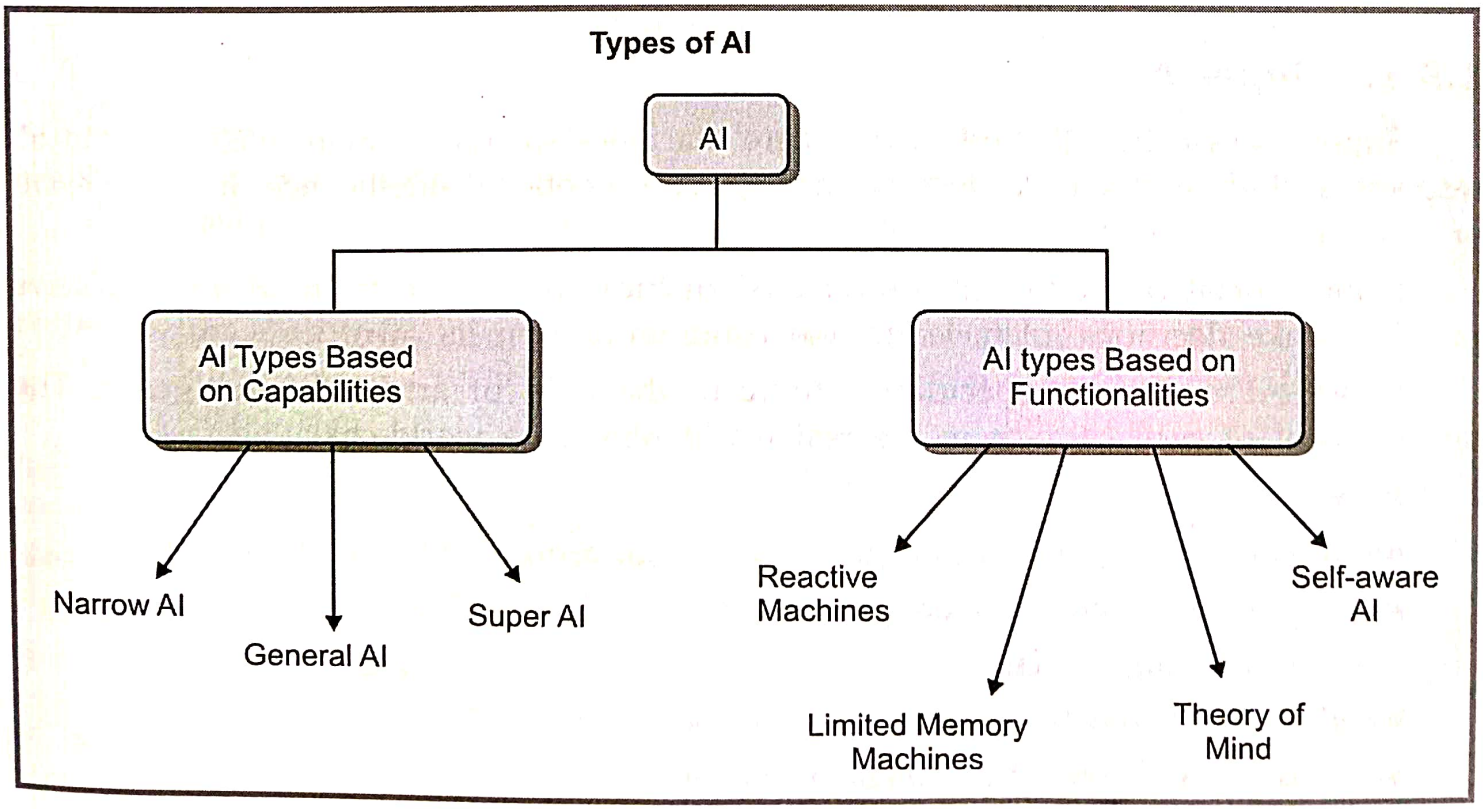
(d) Multiple Instruction Multiple Data (MIMD) Machines

In these machines, numeric computations occupy a substantial chunk of the processing time, followed by symbolic processing.

1.9 TYPES OF AI

1.9.1 CLASSIFICATION BASED ON CAPABILITIES

AI systems are often categorized by their level of capability in mimicking human intelligence. This classification includes three types:



1.9.1.1 Narrow AI (Weak AI)

Narrow AI is designed to perform specific tasks with high efficiency. These systems operate under a limited set of constraints and cannot function beyond their predefined capabilities. The machine intelligence that we see all around us today is an example of narrow AI.

Virtual assistants like Apple's Siri and Alexa, Netflix recommendations, ecommerce purchase suggestions, self-driving cars, chess-playing programs, and speech and picture recognition are all examples of narrow AI applications.

Advantages:

- High precision in specific tasks.
- Easier to develop and deploy.

Limitations:

- Lack of general intelligence.
- Cannot adapt to new tasks without reprogramming.

1.9.1.2 General AI (Strong AI)

General AI refers to systems capable of performing any intellectual task that a human can do. These systems exhibit reasoning, problem-solving, and learning across a wide range of domains.

As systems with general AI are still under research, and it will take lots of efforts and time to develop such systems. The worldwide researchers are now focused on developing machines with General AI. It's potential applications may be in fully autonomous robots and fully AI capable scientific research.

1.9.1.3 Super AI

Super AI is a hypothetical stage where machines surpass human intelligence in all respects, including creativity, decision-making, and emotional intelligence. It is the result of general AI.

Some crucial properties of a strong AI include the ability to think, reason, solve puzzles, make decisions, plan, learn, and communicate on its own.

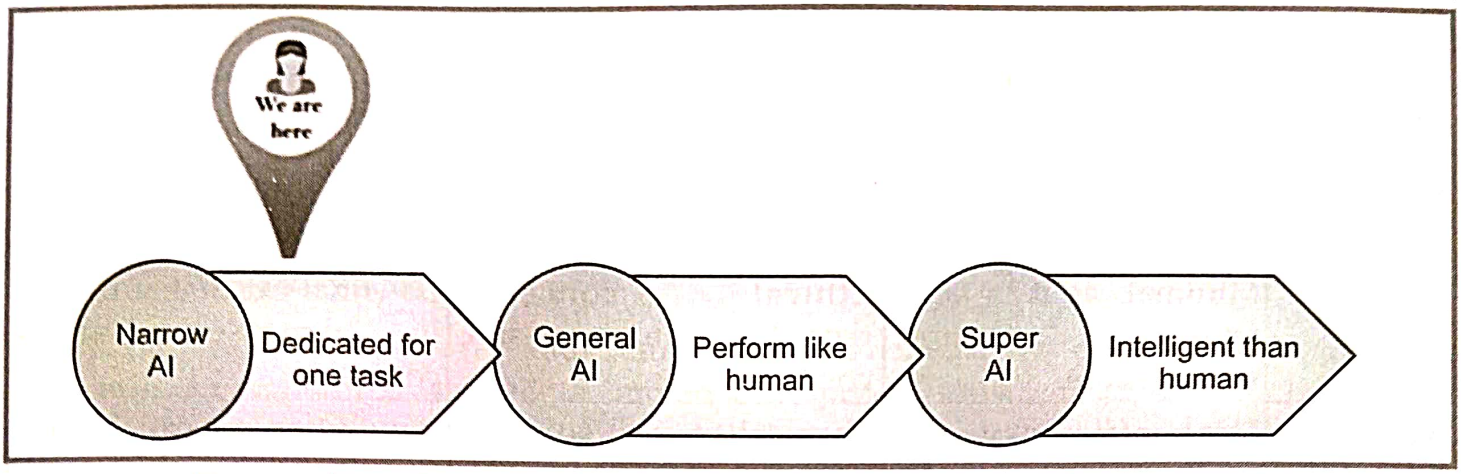
Super AI remains a speculative topic in the field of artificial intelligence. The development of such systems in the real world remains a world-changing task.

Potential applications:

- Revolutionizing fields like healthcare, engineering, and education.
- Addressing complex global challenges like climate change.

Concerns of super AI:

- Ethical dilemmas regarding control and misuse.
- Risk of AI surpassing human oversight.



Comparison between Narrow AI, General AI, and Super AI

Feature	Narrow AI (Weak AI)	General AI (Strong AI)	Super AI
Definition	AI designed for specific tasks and limited functionality.	AI with the ability to understand, learn, and perform any intellectual task like a human.	Hypothetical AI surpassing human intelligence in all aspects.
Scope	Task-specific; cannot function beyond its programming.	Broad and versatile, capable of solving problems across domains.	All-encompassing; exceeds human abilities in reasoning, creativity, and decision-making.
Examples	Voice assistants like Siri and Alexa.	Not yet fully developed but envisioned as future AI systems.	Hypothetical future AI, such as advanced autonomous decision-making entities.
Learning Ability	Limited to the predefined dataset and tasks.	Can learn from experience and adapt to new tasks autonomously.	Capable of self-learning and improving without human intervention.
Cognitive Abilities	Lacks human-like reasoning or emotional understanding.	Mimics human cognitive abilities such as reasoning, planning, and problem-solving.	Surpasses human cognitive abilities, including emotions and intuition.
Real-World Examples	Spam filters, recommendation systems, and autonomous vehicles.	Currently theoretical; often depicted in AI research or fiction.	Not yet realized; discussed in speculative AI debates and ethics.
Development Status	Fully operational and widely used in various industries.	Under active research and experimentation.	Speculative; theoretical discussions dominate this category.

Feature	Narrow AI (Weak AI)	General AI (Strong AI)	Super AI
Decision-Making	Limited to the predefined algorithms and data patterns.	Capable of independent decision-making based on reasoning.	Possesses superhuman decision-making and problem-solving skills.
Risks	Minimal, as it is task-specific and manageable.	Ethical concerns regarding autonomy and control.	Potential existential risks if it surpasses human control.
Advan-tages	High efficiency and accuracy for specific tasks.	Versatility in solving complex, multi-domain problems.	Revolutionary capabilities to solve global challenges.
Challenges	Cannot adapt to unforeseen scenarios.	equires significant advancements in AI research.	Raises profound ethical and existential questions.

The transition from Narrow AI to General AI and potentially to Super AI involves significant technical, ethical, and societal challenges, making this an area of both excitement and caution in AI development.

- **Narrow AI** is practical, currently in use, and effective for specific applications.
- **General AI** represents the aspiration to create machines that think and act like humans across all domains.
- **Super AI** remains hypothetical but inspires both hope for revolutionary advancements and concerns about its implications.

1.9.2. CLASSIFICATION BASED ON FUNCTIONALITIES

AI systems can also be categorized by their functional abilities and how they interact with the environment. The system's capacity for information processing, learning, and decision-making is the main focus of this classification. It offers information on the operational level of AI, from simple task-performing systems to ones with sophisticated reasoning and emotional intelligence.

1.9.2.1 Reactive Machines

Purely reactive machines are the most basic types of Artificial Intelligence. Reactive machines operate solely based on the present data without any memory or past experience. These machines only focus on current scenarios and react on it as per possible best action. Reactive AI stems from statistical math and can analyze vast amounts of data to produce a seemingly intelligent output. They are quick and efficient for specific tasks but cannot improve over time or learn from history.

Examples:

- IBM's Deep Blue: By examining the pieces on the board and forecasting the likely outcomes of each move, IBM's chess-playing supercomputer AI defeated chess grandmaster Garry Kasparov in the late 1990s.
- Google's AlphaGo

1.9.2.2 Limited Memory

These systems can retain data temporarily and use it for decision-making. These machines can use stored data for a limited time period only.

Though AI with limited memory can use historical data for a limited length of time, it is unable to store that data in a library of prior experiences for later use. Over time, Limited Memory AI might perform better as it is trained on more data.

Limited Memory AI shows better performance than reactive machines and are suitable for dynamic environments. However, they have limited capability to handle long-term knowledge.

Examples:

- To forecast the next word, phrase, or visual element in the content they are creating, generative AI products like ChatGPT, Bard, and DeepAI rely on limited memory AI capabilities.
- Chatbots and virtual assistants: Siri, Alexa, Google Assistant, Cortana, and IBM Watson Assistant use limited memory artificial intelligence (AI) and natural language processing (NLP) to comprehend queries and requests, respond appropriately, and create responses.
- Self-driving cars: AI with limited memory helps autonomous cars comprehend their surroundings in real time and decide whether to brake, accelerate, turn, and other maneuvers.

1.9.2.3 Theory of Mind AI

Theory of Mind AI is a functional class of AI that falls underneath the General AI. This theoretical type of AI understands emotions, beliefs, and intentions, enabling it to interact more effectively with humans. In theory, this would allow the AI to simulate human-like relationships.

Although these kinds of AI devices have not yet been created, academics are working hard to make improvements in this area. Theory of Mind AI would tailor its interactions with people according to their particular emotional needs and intentions since it could deduce human motivations and reasoning. In addition, Theory of Mind AI would be able to comprehend and interpret essays and artwork, something that current generative AI tools cannot do. To date, Emotion AI is unable to understand and respond to human feelings.

Potential applications can be :

- AI therapists or companions.
- Advanced personal assistants.

1.9.2.4 Self-Aware AI

Self-awareness Artificial Intelligence is the way of the future. Self-Aware AI is purely theoretical, just like theory of mind AI. If it were ever accomplished, it would be able to comprehend human emotions and thoughts as well as its own internal conditions and characteristics. It would also have a unique set of needs, feelings, and convictions. It can be highly beneficial to develop highly advanced applications in every field.

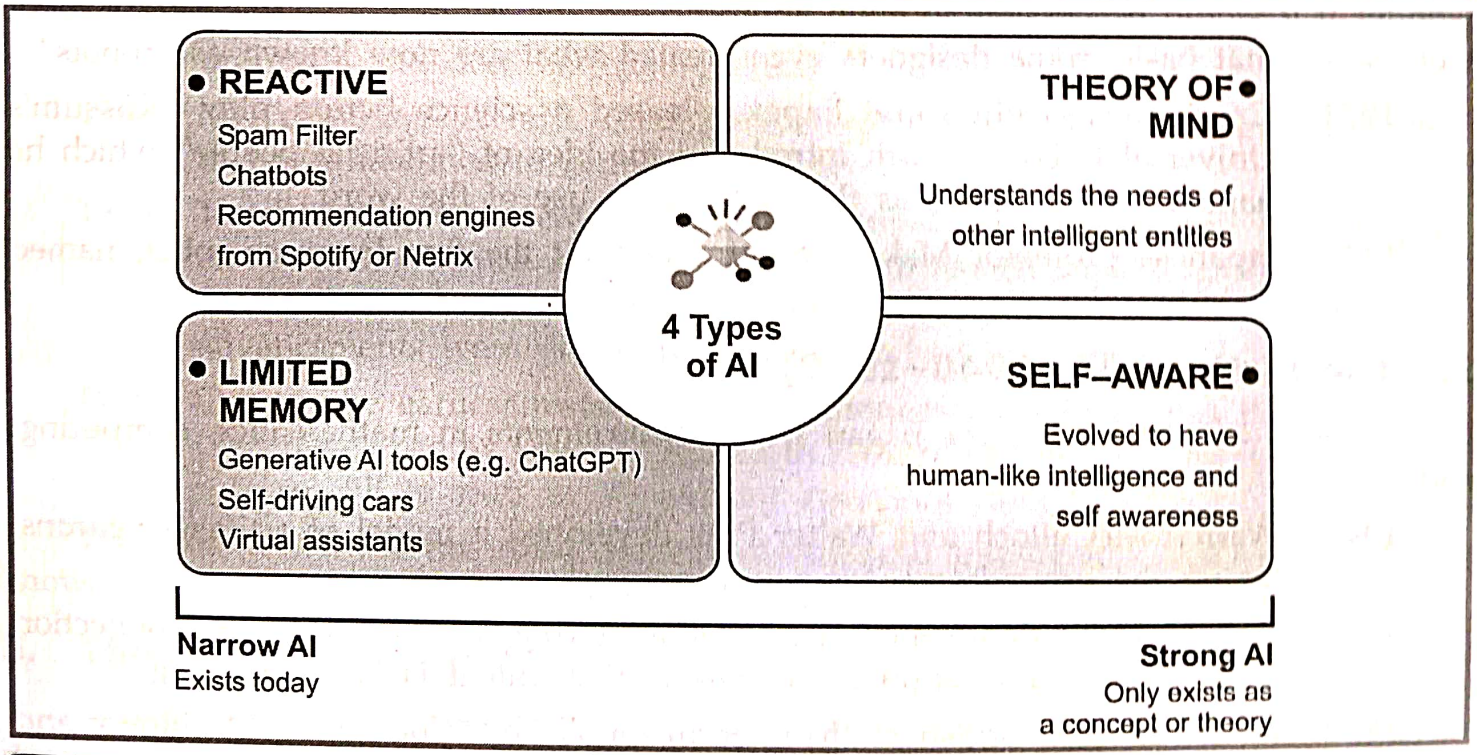
There can be some ethical considerations in Self Aware AI about machine rights and responsibilities.

The human mind will not be as intelligent as these machines. Self-Awareness AI is a theoretical idea that does not yet exist in reality.

The above mentioned four categories represent the progressive evolution of AI capabilities, from simple task-based systems (Reactive Machines) to hypothetical, advanced AI capable of self-awareness. While Reactive Machines and Limited Memory AI are commonly found today, Theory of Mind and Self-Aware AI remain speculative and under development, reflecting the potential future of AI's integration into human society. The following table summarizes the key differences between **Reactive Machines, Limited Memory AI, Theory of Mind AI, and Self-Aware AI:**

Feature	Reactive Machines	Limited Memory AI	Theory of Mind AI	Self-Aware AI
Definition	AI that operates based solely on present inputs, without memory or ability to learn.	AI that retains short-term memory to enhance decision-making.	Hypothetical AI capable of understanding emotions, beliefs, and intentions.	Hypothetical AI with self-consciousness and awareness of its own existence.
Memory Capability	No memory; operates only with current data.	Short-term memory; can store and use past information temporarily.	Requires memory to understand others' mental states.	Full memory and self-awareness to understand its existence.
Learning Ability	No learning; fixed programming.	Limited learning from recent data or experiences.	Advanced learning and reasoning about human emotions and interactions.	Fully autonomous learning, reasoning, and self-reflection.
Interaction	Responds directly to stimuli without adapting.	Adapts to situations based on recent observations.	Interacts effectively by understanding human emotions and beliefs.	Self-reflective and capable of complex social interactions.

Feature	Reactive Machines	Limited Memory AI	Theory of Mind AI	Self-Aware AI
Examples	• IBM's Deep Blue (Chess-playing computer). • spam filters.	• Autonomous vehicles using sensor data. • Traffic prediction, speech recognition.	• Hypothetical AI therapists or advanced assistants. • AI companions, emotional interaction systems.	• Currently non-existent, but theorized in advanced robotics. • Advanced robotics with human-like self-awareness.
Complexity	Low; performs predefined tasks efficiently.	Moderate; handles dynamic environments better.	High; capable of complex social interactions.	Very high; surpasses human intelligence in awareness and reasoning.
Current Status	Widely implemented.	Common in modern AI systems (e.g., self-driving cars).	Under research; not yet realized.	Purely theoretical at this stage.



1.10 HISTORY OF AI

For researchers, artificial intelligence is neither a new term nor a new technology. You would be surprised to learn how old this technology is. With myths and legends about

automatons and thinking machines, the concept of building robots that replicate human intellect existed even in history. But it wasn't until the first electronic computers were created in the middle of the 20th century that their full potential was examined. The speed at which AI continues to expand is unprecedented, and to appreciate how we got to this present moment, it's worthwhile to understand how it first began. AI has a long history stretching back to the 1950s, with significant milestones at nearly every decade. Here's a simplified timeline tracing the major milestones of AI:

1. Early Ideas (Before 20th Century)

The idea of intelligent machines has roots in ancient beliefs and myths:

Greek mythology: Early human imagination regarding artificial life was demonstrated by tales of mechanical creatures such as the enormous bronze robot Talos.

Philosophers: René Descartes and other thinkers wondered in the 1600s if machines could "think" like people.

Mathematics: Ada Lovelace hinted to early AI principles when she indicated that Charles Babbage's Analytical Engine, a forerunner to contemporary computers, could do duties beyond computations in the 1800s.

2. Groundwork for AI (1900–1950)

There was a lot of media produced in the early 1900s that focused on the concept of artificial humans. So much so that researchers from many walks of life began to wonder if it would be possible to build an artificial brain. Although the majority of them were somewhat basic, some designers even created what are now known as "robots".

1921: Czech playwright Karel Èapek released a science fiction play "Rossum's Universal Robots" which introduced the idea of "artificial people" which he named robots. This was the first known use of the word.

1929: Japanese professor Makoto Nishimura built the first Japanese robot, named Gakutensoku.

3. The Birth of AI (1940–1950)

AI as a scientific discipline began with advancements in mathematics, computing, and logic:

1943: Warren McCulloch and Walter Pitts developed a model of artificial neurons, laying the groundwork for neural networks.

1949: Donald Hebb demonstrated an updating rule for modifying the connection strength between neurons. His rule is now called Hebbian learning.

1950: Alan Turing, a British mathematician, publishes "Computing Machinery and Intelligence" in which he proposed the Turing Test, a way to determine if a machine can exhibit human-like intelligence.

1951: Marvin Minsky and Dean Edmonds created the initial artificial neural network (ANN) named SNARC. They utilized 3,000 vacuum tubes to mimic a network of 40 neurons.

4. Early AI Research (1950–1970)

This era focused on programming computers to solve simple problems. Programs like Logic Theorist and General Problem Solver demonstrated early reasoning abilities and the first chatbot, ELIZA, simulated human conversation but still the progress was slow due to limited computing power and understanding of AI's complexities in this period.

1956: Dartmouth College mathematics professor John McCarthy invited a small group of researchers from various disciplines to participate in a summer-long workshop focused on investigating the possibility of "thinking machines." The term "Artificial Intelligence" was coined at the Dartmouth Conference, marking the official start of AI as a field of study.

1958: John McCarthy developed the LISP language. LISP stands for List Processing and was soon adopted as the language of choice among most AI developers.

1959: Arthur Samuel while giving a speech about teaching machines to play chess more effectively than the people who designed them, coined the phrase "machine learning".

1961: 'Unimate', the first industrial robot, began operating on an assembly line at General Motors in New Jersey. His duties included welding parts on automobiles and moving die castings, tasks that were considered too risky for humans to perform.

1965: Edward Feigenbaum and Joshua Lederberg developed the first "expert system," a type of artificial intelligence designed to mimic human experts' thought processes and decision-making skills.

1966: MIT computer scientist Joseph Weizenbaum created, ELIZA, the first chatbot that was intended to simulate therapy.

1973: The British government significantly decreased its support and financing for AI research when an applied mathematician named James Lighthill presented a report to the British Science Council highlighting the fact that scientific advancements were not as remarkable as had been promised.

1979: One of the earliest instances of an autonomous vehicle was the Standford Cart, developed by James L. Adams in 1961. It was able to successfully explore a room full of seats in '79 without assistance from humans.

The Association for the Advancement of Artificial Intelligence (AAAI) , formerly known as the American Association of Artificial Intelligence, was established in 1979.

5. The AI Winters (1970 -1980)

AI experienced periods of decline, known as "AI Winters," when funding and interest dropped due to unmet expectations. AI winters is known as a tough period for artificial intelligence (AI). Reasons for the setbacks can be contributed to the over promising results that couldn't be delivered and high costs and limited success in real-world applications. This period is often referred to the gap between AI expectations and the technology's shortcomings.

6. Revival of AI (1980–1990)

AI research regained momentum with new approaches and tools like expert systems, Neural Networks, advancements in hardware, etc.

1980: AI returned with "Expert System" after its hibernation. Expert systems were designed to mimic a human expert's decision-making process.

1981: Danny Hillis developed parallel computers with an architecture similar to modern GPUs that were specifically designed for artificial intelligence and other computational tasks.

1986: Ernst Dickmanns and his team at Bundeswehr University of Munich, invented the first self-driving car. It could drive up to 55 mph on roads that didn't have other obstacles or human drivers.

1987: Commercial launch of Alacrity by Alactrious Inc. Alacrity was the first strategy managerial advisory system, and used a complex expert system with 3,000+ rules.

7. Intelligent Agent development (1993–2011)

Despite the lack of funding during the second AI Winter (1987-1993) , the early 90s showed some impressive strides forward in AI research. AI advanced significantly between 1993 and 2011, especially in the creation of intelligent computer programs. In this period, the focus of AI experts moved from trying to equalize human intelligence to creating clever, useful software for specific jobs. This era also introduced AI into everyday life via innovations such as the first Roomba and the first commercially-available speech recognition software on Windows computers.

The following are some significant events from this period:

1997: IBM's Deep Blue beat the world chess champion, Gary Kasparov, in a highly-publicized match, becoming the first program to beat a human chess champion.

1997: Windows released a speech recognition software (developed by Dragon Systems).

2000: Kismet – a 'social robot' was created in MIT's Artificial Intelligence Laboratory and helmed by Dr. Cynthia Breazeal, that contained sensors, a microphone, and programming that outlined "human emotion processes". It could simulate human emotions with its face, which included eyes, eyebrows, ears, and a mouth.

2002: AI entered the home in the form of Roomba, a vacuum cleaner.

2003: NASA landed two rovers onto Mars (Spirit and Opportunity) and they navigated the surface of the planet without human intervention.

2006: Companies such as Twitter, Facebook, and Netflix started utilizing AI as a part of their advertising and user experience (UX) algorithms.

2010: Microsoft launched the Xbox 360 Kinect, the first gaming hardware designed to track body movement and translate it into gaming directions.

2011: In a televised round of Jeopardy, an IBM Watson, an NLP computer designed to respond to questions, defeated two previous winners.

2011: Apple released Siri, the first popular voice-activated personal virtual assistant.

8. The Big Boom in AI (2012–Present)

With the advent of large data and sophisticated algorithms, AI's capabilities and popularity skyrocketed in the era. This leads us to the most current advancements in artificial intelligence. Common place AI tools like search engines and virtual assistants have become more prevalent. The combination of Big data, deep learning, widespread data application, and the continuous pursuit of artificial general intelligence (AGI) is responsible for these accomplishments. Some of the notable occurrences from this time frame are as follows:

2012: Two researchers from Google (Jeff Dean and Andrew Ng) trained a deep neural network to recognize cats by showing it unlabeled images and no background information.

2012: Google's released "Google Now" function for Android apps, which may forecast information for the user.

2013: DeepMind introduced deep reinforcement learning, a CNN that outperformed human game professionals by learning new abilities through repeated practice and prizes.

2014: Variational Autoencoders (VAEs) were developed by Diederik Kingma and Max Welling to generate text, images, and videos.

2014: Facebook developed the DeepFace deep learning facial recognition technology, which can recognize human faces in digital photos with an accuracy that is almost on par with human talents.

2016: In Seoul, South Korea, DeepMind's AlphaGo defeated the renowned Go expert Lee Sedol, bringing back memories of Kasparov's chess encounter versus Deep Blue over twenty years earlier.

2016: Sophia, the first "robot citizen," was a humanoid robot built by Hanson Robotics. She was the first to have a realistic human look, the capacity to see and mimic emotions, and the ability to speak.

2018: The IBM "Project Debater" performed incredibly well while debating difficult subjects with two expert debaters.

2018: In a demonstration of Google's AI program "Duplex," the woman on the other end of the line was unaware that she was speaking to a virtual assistant that had scheduled hair salon appointments on call.

2019: In the video game StarCraft 2, Google's AlphaStar defeated all but .2% of human players to attain Grandmaster.

2020: GPT-3, a model that leverages Deep Learning to generate code, poetry, and other language and writing tasks, began beta testing by OpenAI. Although it is not the first of its kind, it is the first to produce content that is nearly identical to human-made stuff.

2021: DALL-E was created by OpenAI, which advances AI's comprehension of the visual world by processing and comprehending images sufficiently to provide precise captions.

2022: Open AI launched ChatGPT, offering a chat-oriented interface to its GPT-3.5 LLM.

2023: Google released their GPT chatbot Bard and Microsoft incorporate ChatGPT into its search engine Bing, OpenAI also launched GPT-4, which once again improved on the strength of its predecessor. In addition to GPT-4 can now do an ever-widening range of tasks, including passing the bar exam, and produce significantly more complex and imaginative answers.

2024: This year has been proved to be a crucial year for AI's future as scientists and businesses work to determine the most practical ways to incorporate this technological advancement into our daily lives through Generative AI break throughs.

9. The Future of AI

AI continues to advance at an unprecedented pace, with promising developments in achieving Artificial General Intelligence (AGI) and tackling ethical concerns. AI is revolutionizing the fields like healthcare, education, and climate science.

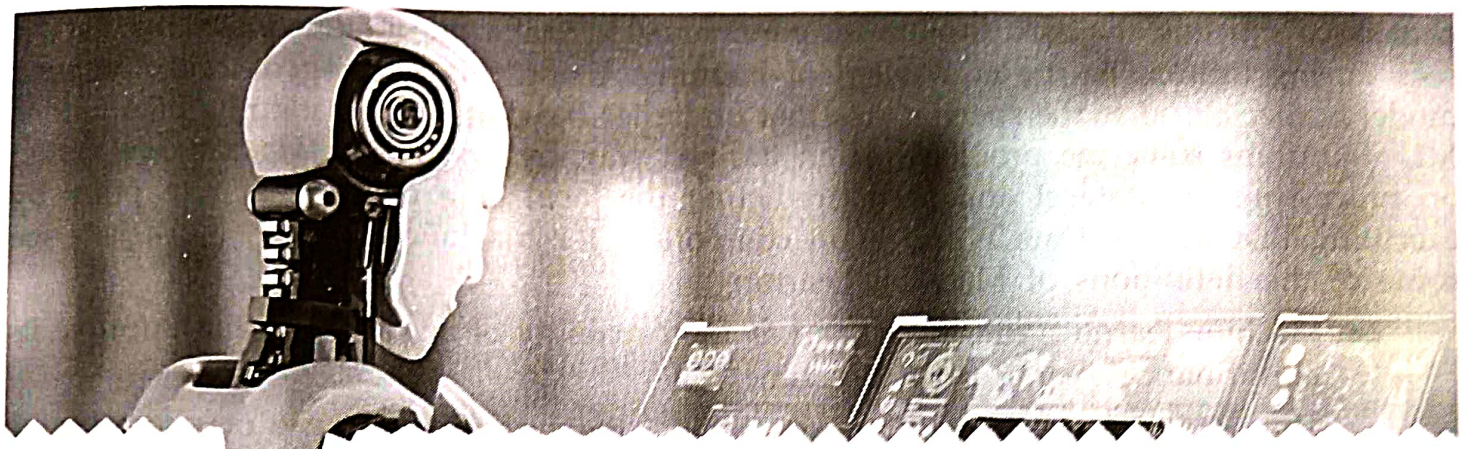
From ancient myths to modern breakthroughs, AI has come a long way. It is now an integral part of our lives, with the potential to shape the future of humanity. The history of AI highlights cycles of progress and setbacks. From Turing's foundational ideas to modern AI's widespread applications, the journey reflects humanity's quest to create intelligent machines. While challenges remain, the future of AI promises even greater possibilities. As we continue to explore and refine AI, its history reminds us of the power of human creativity and innovation.

Although we can never fully predict the future, we can make educated guesses based on the opinions of numerous leading experts. We can anticipate that businesses of all sizes will continue to adopt AI, the workforce will change as more automation creates and eliminates jobs in equal measure, and there will be more robotics, autonomous vehicles, and much more.

QUESTIONS

1. What is Artificial Intelligence ?
2. Explain the key goals of AI ?
3. Discuss the characteristics of AI.
4. How can AI be categorised ?
5. Explain the major developments of AI in history.
6. Whether AI is Boon or Bane. Comment.
7. Differentiate between :
   (a) Narrow AI, General AI and Super AI.
   (b) Reactive machines, Limited theory, Theory of mind AI and self aware AI.

☆☆☆☆☆



CHAPTER

2 INTRODUCTION TO KEY TECHNOLOGIES OF AI

CHAPTER AT A GLANCE

2.1 Machine Learning	2.4 Computer Vision
2.2 Deep Learning	2.5 Robotics
2.3 Natural Language Processing (NLP)	

Artificial Intelligence (AI) comprises a range of key technologies, each playing a pivotal role in transforming industries and enabling innovative solutions. These technologies include **Machine Learning (ML), Natural Language Processing (NLP), Computer Vision, Robotics,** and **Expert Systems,** among others.

2.1 MACHINE LEARNING

The idea of an algorithm is fundamental to computer programming. An algorithm is a set of instructions that must be followed in order to convert an input into an output. Since the invention of computers, we have developed algorithms for a wide range of tasks, and as a result, we use computers for a variety of purposes today. Digital technology has emerged as the primary method for storing, processing, and transmitting information, and they are now an essential part of our daily lives on both a professional and social level.

However, despite decades of research, we still lack an algorithm for some tasks. Humans are capable of performing some of these tasks with ease and without even realizing it. We are able to play chess, drive a car, converse in a foreign language, and identify people from photographs. We can also navigate a crowded space without hitting with people or objects. Learning to do these kinds of tasks is the goal of machine learning. Arthur Samuel first used the term "machine learning" in 1959.

Our strategy is to begin with a very general model that has a lot of parameters. Depending on how the parameters are set, this general model can perform a wide range

of tasks. Adjusting the values of those parameters to make the model match the data it sees during training is the equivalent of learning. Using this training data, the general model is specialized to the specific task that underlies the data by adjusting its parameters. ML algorithms solve the problems in an indirect way, by first generating a model based on processing the dataset and then predicting the label of a new input data point by executing that model. Data is the main component for machine learning. Now, let's see some of the definitions of Machine Learning;

- Machine learning can be defined as the capability of an AI system to improve its performance over a period of time.

  This, of course, assumes the capability of the system to acquire new knowledge and skills, as well as its capability to recognize the existing knowledge based on the newly acquired knowledge.
- Machine learning is a branch of artificial intelligence that provides systems with the ability to learn from experience without being programmed explicitly.
- Machine learning (ML) is a field of computer science that studies algorithms and techniques for automating solutions to complex problems that are hard to program using conventional programing methods.

The development of computer programs that can access data and learn from it on their own is the focus of machine learning. Data or observations, such as instruction, first-hand experience, or examples, are the starting point for the learning process. Patterns are then identified from the data and used to inform future predictions.

2.1.1 GENERIC WORKING OF MACHINE LEARNING MODELS:

When given new data, a machine learning system creates prediction models, learns from it, and forecasts its results. The quantity of data influences the accuracy of the predicted output by helping in the development of a more accurate model.

Let's say we have to make predictions about a complicated problem. We only need to supply the data to generic algorithms, which generate the logic based on the data and forecast the result, rather than writing code. Machine learning has altered our understanding of the problem. The block diagram in figre 2.1 illustrates how the machine learning algorithm works:

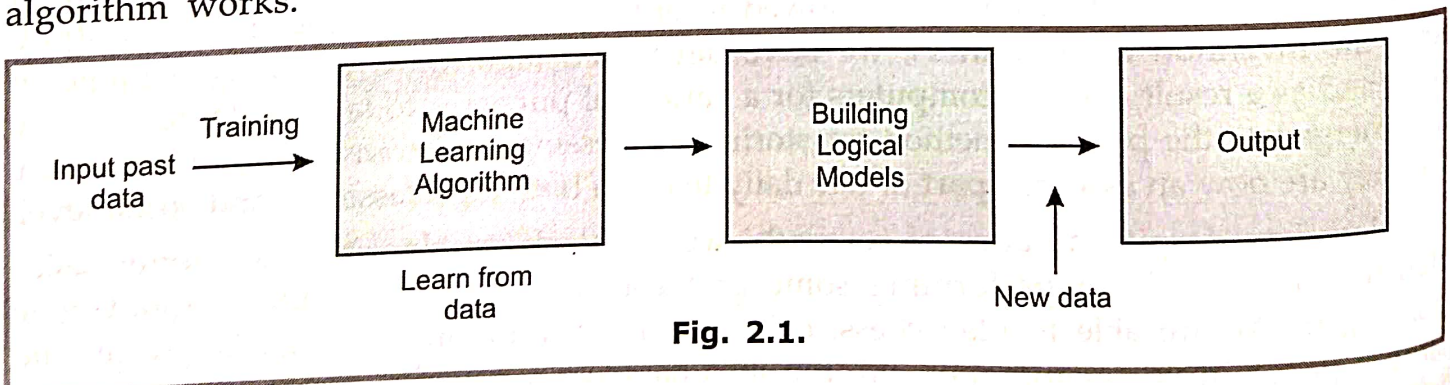


Fig. 2.1.

2.1.2 NEED OF MACHINE LEARNING

- **Handling the Big Data Challenge**: Conventional approaches are being overtaken by the exponential growth of data from multiple sources, such as social media, Internet of Things devices, and business transactions. ML algorithms can process,

analyze, and derive insights from vast datasets efficiently, enabling data-driven decision-making.

- **Automating the Complex Processes:** Some tasks, such as image recognition or natural language processing, are too complex for traditional programming. ML automates these tasks by learning patterns and making predictions, reducing the need for manual intervention.
- **Enhancing Personalization:** Customers demand personalized experiences, which are hard to deliver at scale using manual methods. ML algorithms analyze user behavior to create tailored recommendations, such as Netflix suggesting shows or Amazon recommending products.
- **Solving Complex Problems:** Challenges like climate prediction, disease diagnosis, and fraud detection are intricate and data-intensive. ML excels in identifying patterns and relationships in complex datasets, enabling breakthroughs in these areas.
- **Dynamic Adaptation:** Environments change rapidly, requiring systems to evolve. ML systems learn and adapt continuously, making them suitable for dynamic applications like autonomous vehicles and cyber security.
- **Improving Decision-Making:** Businesses and industries require accurate, real-time decisions in dynamic environments. ML models provide predictive insights and recommendations, enhancing decision-making in fields like finance, healthcare, and supply chain management.

2.1.3 CHARACTERISTICS OF MACHINE LEARNING

Machine learning has a number of features, including:

1. Generate extremely accurate predictions based on test data (discovering the underlying "truth" is not the aim).
2. Techniques should be "off-the-shelf," general-purpose, and completely automated; in reality, though, it is essential to incorporate previous, human knowledge.
3. Theory and practice should interact richly
4. The focus is given on techniques that manage big data sets.
5. Enable computers to learn on their own without human assistance and adapt as necessary.
6. Machine learning allows us to analyze vast amounts of data.

2.1.4 RELATION WITH AI

- Compared to machine learning (ML), AI is a far more expansive field of study. While machine learning (ML) is primarily about creating machines that can learn to perform tasks, AI is all about making machines intelligent using multiple approaches.
- While intelligence is difficult to define, machine learning is undoubtedly a branch of artificial intelligence.

- While AI encompasses all technologies that enable machines to perform tasks that typically require human intelligence, ML focuses specifically on enabling systems to learn from data and improve their performance without explicit programming.
- While AI sets the goal of intelligent behavior, ML provides the methods to achieve this goal by leveraging data. ML techniques, such as supervised, unsupervised, and reinforcement learning, underpin many AI applications. Advanced AI systems like IBM Watson, Google Assistant, and autonomous vehicles rely heavily on ML to function effectively.

2.1.5 CLASSIFICATION OF MACHINE LEARNING

Machine Learning algorithms derive their strength from their ability to learn from the available data. Before going further let's formally understand the process of 'learning'. At a broader level, machine learning can be classified into three types:

2.1.5.1 Supervised learning

Supervised learning is concerned with forecasting the future events statistically based on the fed historical data. In supervised learning (SL), the machine is given a dataset (i.e., a set of data points), along with the right answers to a question corresponding to the data points. A large collection of data points with answers (a labelled dataset) is given to the learning algorithm. To find the answer, the algorithm must discover the salient features of every data point in the dataset. Therefore, the algorithm (machine) should be able to predict the outcome or correct answer the next time a new data point is given to it based on the key characteristics.

Let us explain it with an illustration:

A child learns to recognize objects, like cats, by being exposed to various examples (pictures and real-life instances) and receiving correct labels (e.g., "cats"). Occasionally, the child makes mistakes, such as identifying a wolf as a cat, and corrections from his parents, family, etc. help him to refine the understanding over time.

In the context of **machine learning**, supervised learning involves a machine being provided with labeled data, such as images with corresponding labels (dog, cat, car, etc.) or stock market data with outcomes. The machine analyzes the data to learn patterns and relationships, enabling it to make predictions on new, unseen data, like identifying an image or predicting stock behavior.

The diagram below shows the diagrammatic representation of Supervised learning. As shown in figure 2.2 we have a dataset of images of carrots, tomato and bell peppers. Every image in the dataset is labeled, means we know which image represents a tomato or carrot, or which image represents a bell pepper. We give all the images and their corresponding labels to the model/algorithm. Then, it should be able to make the difference on its own. If everything goes very well, when showing a new image to the model without having its label, it should be able to tell us the name of the vegetables among the three (if the model learned well).

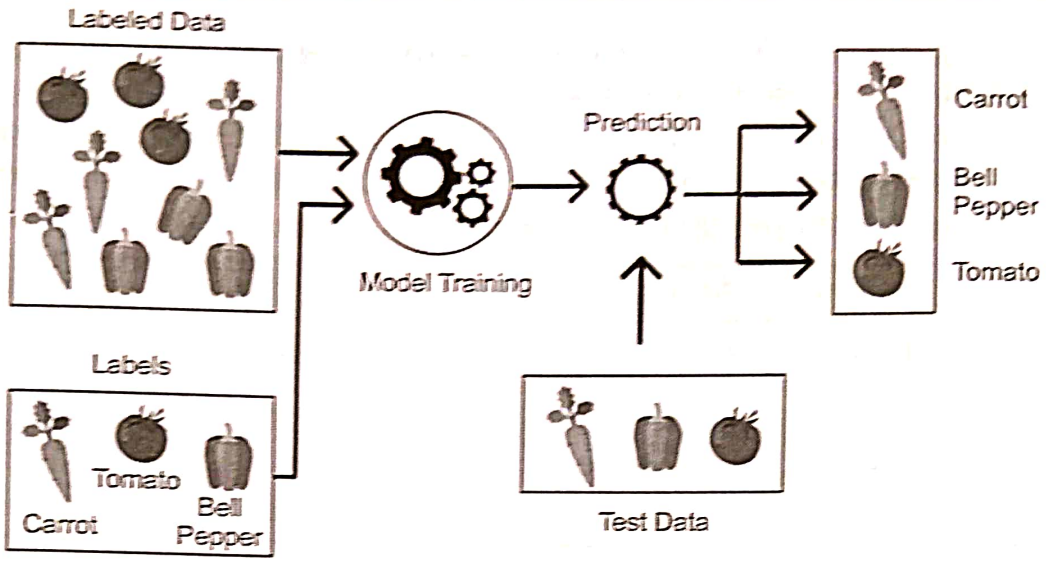


Fig. 2.2. Diagrammatic representation of supervised learning.

Consider the Table 2.1 dataset, an example of how supervised learning can be applied in medical diagnostics. Each row in the dataset represents a patient's health metrics, which serve as the **input features**. These features are used to predict a corresponding **label** ("healthy" or "sick").

Table 2.1: Dataset of Patient's Helath Matrices

Blood Pressure	Cholesterol Level	Heart Rate	Label (Disease)
120/80	180	72	Healthy
140/90	240	90	Sick
130/90	160	78	Healthy
145/100	200	85	Sick

- **Training Data:** A collection of rows like this with known features and corresponding labels is provided to the model.
- **Model Learning:** The machine learning algorithm analyzes the patterns in the input features (e.g., how high cholesterol and elevated blood pressure correlate with the disease).
- **Prediction:** Once trained, the model can predict whether a new patient (with unseen data) has the disease based on their health metrics.

Sorting spam into a different folder from your inbox, Recommending your next purchase, smart reply in emails, etc. are just few examples of the many real-world issues that supervised learning helps organizations solve at scale.

Supervised learning can be grouped further in two categories of algorithms:

2.1.5.1.1 Classification: Classification is the process of grouping the output into separate classes based on one or more input variables. Classification is used when the output variable's value is discrete or categorical, like "No" or "Yes" and 0 or 1 and so on, or when the algorithm tries to classify input variables into two different classes, like email, or when the algorithm tries to classify input variables into more than two classes, like

handwritten character recognition, where classes may range from 0 to 9. The Classification can be Binary (e.g., Yes/No), Multiclass (e.g., Cat, Dog, Bird) or Multilabel (Instances can belong to multiple categories simultaneously).

Consider Table 2.2 dataset where the task is to predict whether a customer will default on a loan (Yes/No) based on the features in the table below:

Table 2.2: Example Dataset for Classification Algorithms

ID	Age	Income	Marital Status	Owns House	Defaulted Loan
1	25	30,000	Single	No	No
2	40	75,000	Married	Yes	No
3	35	50,000	Single	Yes	Yes
4	50	100,000	Married	Yes	No

In the given dataset, input are the features (e.g., Age, Income) while the output is the discrete label (e.g., Defaulted Loan: Yes/No). The objective is to learn a mapping from input features to the target label using a labeled dataset. Some of the majorly used techniques for classification are:

- **Logistic Regression:** A statistical method that models the probability of a class using a logistic function. It works well for linearly separable data.
- **Support Vector Machines (SVM) :** Finds the hyperplane that best separates classes in a high-dimensional space. It can handle non-linear data using kernel tricks.
- **Naïve Bayes:** A probabilistic classifier based on Bayes' theorem, assuming independence among features.
- **Classification Trees:** They function similarly to regression trees and are useful for classification problems. Additionally, because of their hierarchical structure, they are scalable, resilient to outliers, and naturally model the non-linear decision boundaries.

2.1.5.1.2.Regression: If there is a relationship between the input and output variables, regression algorithms are employed. It is used in situations where the output variable's value is real or continuous, like in stock price prediction, weather forecasting, and home price estimation.

Table 2.3 is an example dataset for regression, which involves predicting a continuous numerical output based on input features. This example uses housing prices as the target variable.

Table 2.3: Example Dataset for Regression Algorithms

House Size (sqft)	Number of Bedrooms	Age of House (years)	Distance to City Center (miles)	Price ($) (Target)
1200	3	10	5	300,000
1500	4	5	3	400,000
800	2	20	8	200,000
2000	5	2	2	500,000
1000	2	15	7	250,000

This dataset can be used to train regression models to predict the price of house. The goal here is to predict a value as much closer to the actual output value as the model can and then evaluation is done by calculating the error value. The smaller the error the greater the accuracy of the regression model.

Following major techniques are used to train the regression models:

- **Random Forest:** An ensemble of decision trees that improves accuracy and reduces overfitting by averaging multiple tree predictions. For example, predicting crop yields based on soil and weather conditions.
- **Decision Trees:** Builds a tree-like structure to make predictions based on feature splits. For example, predicting energy consumption based on weather data.
- **Linear Regression:** A simple method that assumes a linear relationship between the input features and the target variable. For example, predicting house prices based on square footage.
- **Neural Networks:** Uses deep learning to model complex relationships between features and targets, capable of capturing non-linear patterns. For example, predicting demand for products based on historical sales data.

Table 2.4: Comparison between Regression and Classification

Aspect	Classification	Regression
Definition	Predicts a discrete category or class label.	Predicts a continuous numerical value.
Output	Class labels (e.g., Yes/No, Dog/Cat, 0/1).	Real-valued numbers (e.g., price, temperature, salary).
Type of Problem	Decision-making or categorization tasks.	Prediction or estimation tasks.
Examples	• Spam email detection (Spam/Not Spam) • Diagnosing diseases (Yes/No) • Image classification (Dog/Cat/Car).	• Predicting house prices ($200,000 to $500,000) • Forecasting stock prices over time. • Predicting rainfall in inches.
Algorithms Used	• Logistic Regression • Classification Trees, Naïve Bayes • Support Vector Machines (SVM)	• Linear Regression • Decision Trees, Random Forests • Neural Networks
Output Interpretation	• Probability of belonging to a class (e.g., 80% chance of "Yes").	• Exact numerical value (e.g., house price = $350,000).
Applications	• Medical diagnosis, fraud detection, sentiment analysis.	• Financial forecasting, energy consumption prediction, sales trends.
Key Difference	Deals with assigning data points to specific categories or groups.	Deals with estimating or predicting continuous quantities.

2.1.5.2.Unsupervised learning: Unsupervised learning deals with unlabeled data, means, here we have input data and no corresponding output variable. The goal is to model the underlying distribution or structure of the data and learn more from the data. The name *unsupervised* comes from the fact that there is no teacher and the correct answers are not known. The algorithm is tasked with discovering the underlying data structure on its own. This is the opposite of supervised machine learning. In unsupervised learning, the users do not need to teach/supervise the model. There is no correct output and no supervisor to teach. The algorithm itself learns from the input data & discovers the patterns & information from the data to learn & group the data according to similarities.

As shown in the *Figure 2.3*, suppose we have an image dataset of different types of vegetables. The unsupervised learning algorithm is given that dataset. The algorithm is never seen/trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithms will perform this task by clustering the image dataset into groups according to the similarities between the images.

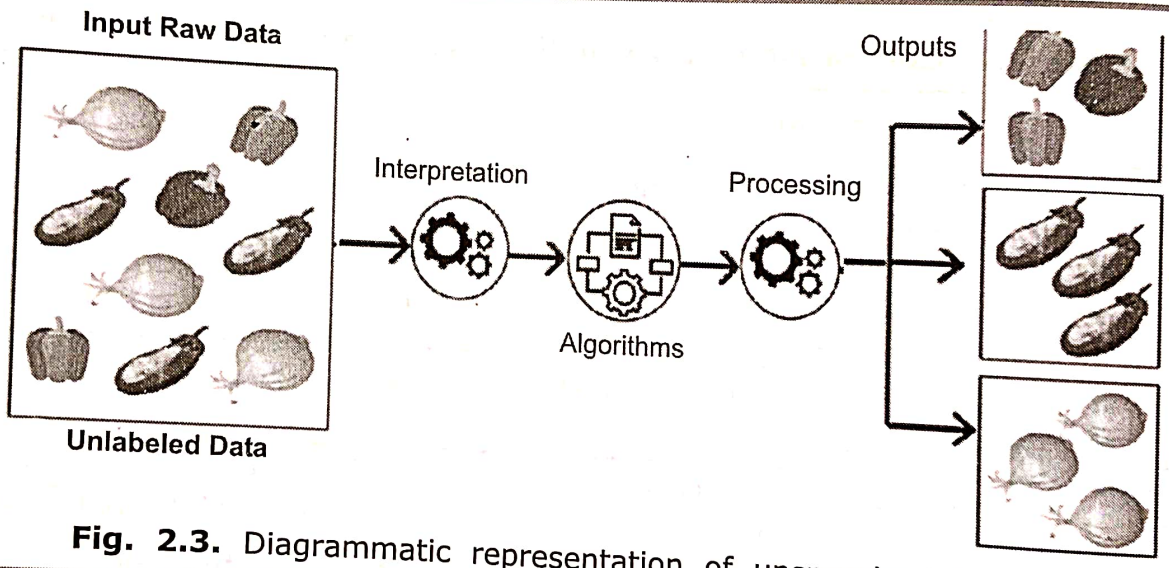


Fig. 2.3. Diagrammatic representation of unspervised learning

Consider the table below, the dataset contains only input features (e.g., age, income, spending score). There are no labels or target outputs. The objective of the unsupervised learning is to identify patterns or group customers based on similarities in the data (e.g., clustering). The clustering will represent groups of customers with similar purchasing behaviors, enabling targeted marketing strategies or business decisions.

Table 2.5: Example Dataset for Unsupervised Learning

Customer ID	Age	Annual Income ($)	Spending Score (1-100)
1	25	50,000	80
2	45	70,000	30
3	35	60,000	60
4	23	30,000	90
5	50	90,000	20

Unsupervised learning algorithms can be classified as:

2.1.5.2.1 Clustering: Clustering is the process of assembling unlabeled data into groups according to shared characteristics. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities. Finding patterns and connections in the data without any prior understanding of its meaning is the aim of clustering.

Some of the common techniques used for clustering can be given as:

- **K-Means Clustering:** Partitions data into k clusters, where each data point belongs to the cluster with the nearest mean. This techniques is simple and efficient for large datasets. It can be efficiently used for Customer segmentation, document clustering, and image compression.
- **Hierarchical Clustering:** Builds a tree-like structure of clusters (dendrogram). It does not require specifying the number of clusters initially. This type of clustering is used in Gene sequence analysis, social network analysis, etc.

2.1.5.2.2 Dimensionality Reduction: Dimensionality reduction is the process of reducing the number of input features (dimensions) in a dataset while retaining as much relevant information as possible. This technique simplifies datasets, making them easier to visualize, process, and analyze, especially when dealing with high-dimensional data.

Dimensionality reduction simplifies datasets by reducing the number of features, making models faster and less resource-intensive while improving computational efficiency. It helps to eliminate irrelevant or redundant features, thereby minimizing noise and focusing on meaningful information. Additionally, it enhances data visualization by transforming high-dimensional data into 2D or 3D spaces, facilitating better interpretation. By addressing the challenges of high-dimensional spaces, where data becomes sparse and traditional algorithms struggle, dimensionality reduction effectively mitigates the curse of dimensionality.

The most common techniques for dimensionality reduction are:

- **Principal Component Analysis (PCA) :** It transforms data into a new coordinate system by identifying the directions (principal components) that capture the most variance in the data. It is a linear technique and is suitable for continuous numerical data. It can be effectively used for image compression, gene expression analysis, finances, etc.
- **Linear Discriminant Analysis (LDA) :** It projects data onto a lower-dimensional space to maximize class separability. Unlike PCA, it uses class labels. It is also a linear technique that combines dimensionality reduction with classification. It can be used for preprocessing step for classification tasks, especially in text or image data.

2.1.5.2.3. Association Rule Learning: An association rule is used to find out the relationships between the data in the large dataset. This method is a rule-based machine learning technique that identifies some extremely helpful relationships between a large data set's parameters. This method is primarily applied to market basket analysis, which aids in a deeper comprehension of the connections among various products.

For instance, based on consumer behavior, retail establishments employ algorithms based on this technique to determine the correlation between the sales of one product and those of another. For example, if a customer purchases milk, he might also purchase bread, eggs, or butter. When properly trained, these models can be used to plan various offers and boost sales.

2.1.5.3 Reinforcement learning: This is machine learning that is based on feedback. Without any labeled data, the machine or agent learns atomically through reinforcement learning. In this case, the agent gains knowledge from its experiences.

The hit-and-trial method is employed by the agent in Reinforcement Learning. The agent receives a reward if they attempt a successful step; if they make a mistake, they receive a penalty. Its objective is to optimize the overall benefit.

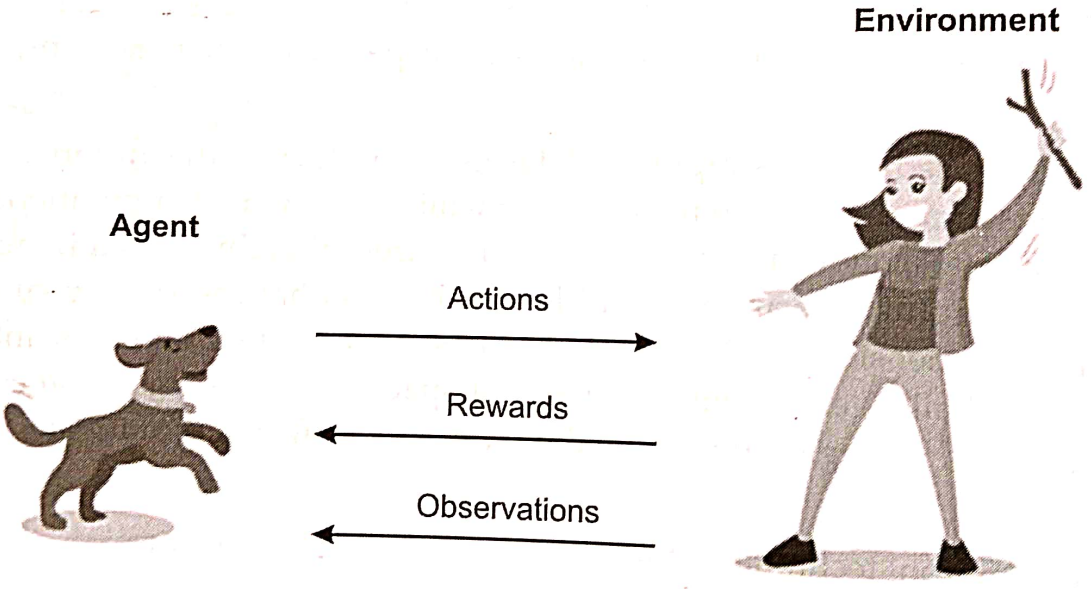


Fig. 2.4. Representation of Reinforcement learning.

Consider the figure 2.4, the example of teaching a dog, is shown in it. We can't have vertal communication with dog because it doesn't understand human language. We create a situation, and the dog responds in various ways. If the dog responds positively, we will give him bread. When the dog is presented with the same situation, it anticipates a higher reward (bread) and repeats the behavior. It's similar to understanding how a dog learns from positive experiences. Simultaneously, the dog is developing the ability to ignore negative experiences.

The table 2.6 below presents a comparison between the three types of discussed Machine learning

Table 2.6. Comparison of Supervised, Unsupervised and Reinforcement Learning

Aspect	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The model is trained on labeled data to predict the output.	The model is trained on unlabeled data to find patterns or structures.	The agent learns by interacting with the environment and receiving feedback through rewards or punishments.
Data Requirement	The model is trained on labeled data to predict the output.	Does not require labeled data; uses only input data.	No labeled data; the agent learns through trial and error.
Goal	Learn a mapping from inputs to outputs (classification or regression).	Find hidden patterns, groupings, or structures in the data (clustering or association).	Maximize cumulative reward by taking actions in an environment.
Example Problems	• Email spam classification • Predicting house prices (regression)	• Customer segmentation • Market basket analysis (association)	• Video game AI (playing chess, Go) • Self-driving cars navigating through traffic
Output	Discrete labels or continuous values.	Grouping of data points (clusters) or learned structures.	A sequence of actions taken by the agent based on its policy.
Algorithms Used	• Linear Regression • Logistic Regression • Support Vector Machines (SVM)	• K-Means Clustering • DBSCAN, Hierarchical Clustering • Principal Component Analysis (PCA)	• Q-Learning • Deep Q-Networks (DQN) • Proximal Policy Optimization (PPO)
Learning Process	The algorithm learns from the labeled data by minimizing an error function.	The algorithm tries to find hidden patterns without any explicit supervision.	The agent learns through interaction, receiving rewards or penalties based on its actions.

Aspect	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Types of Tasks	• Classification (discrete output) • Regression (continuous output)	• Clustering (grouping similar data points) • Dimensionality Reduction (e.g., PCA)	• Decision-making, Policy learning • Multi-agent systems, Games
Feedback	Feedback is provided through the labels of the training data.	No explicit feedback is provided; the model discovers patterns autonomously.	Feedback comes in the form of rewards or penalties after actions are taken.

2.1.6 ADVANTAGES OF ML

There are many advantages of machine learning. These are some of the helpful advantages:

- Often, machine learning is much more accurate than human-crafted rules since it is data-driven.
- Humans are often incapable of expressing what they know (e.g., the rules of English or how to recognize letters), but can easily classify examples.
- Machine learning does not need a human expert or programmer.
- **No human intervention needed or Automation of** Machine learning has a very powerful tool to automate various decision-making tasks.
- **Easily identifies trends and** Machine learning takes large amounts of data and discovers the hidden structures, specific styles and patterns that are very difficult to find by humans.
- **Machine Learning is efficient in handling of any type of Data.**
- **Continuous Improvement:** As machine learning algorithms gain experience from data, they keep improving their accuracy and efficiency.

2.1.7 CHALLENGES OF ML

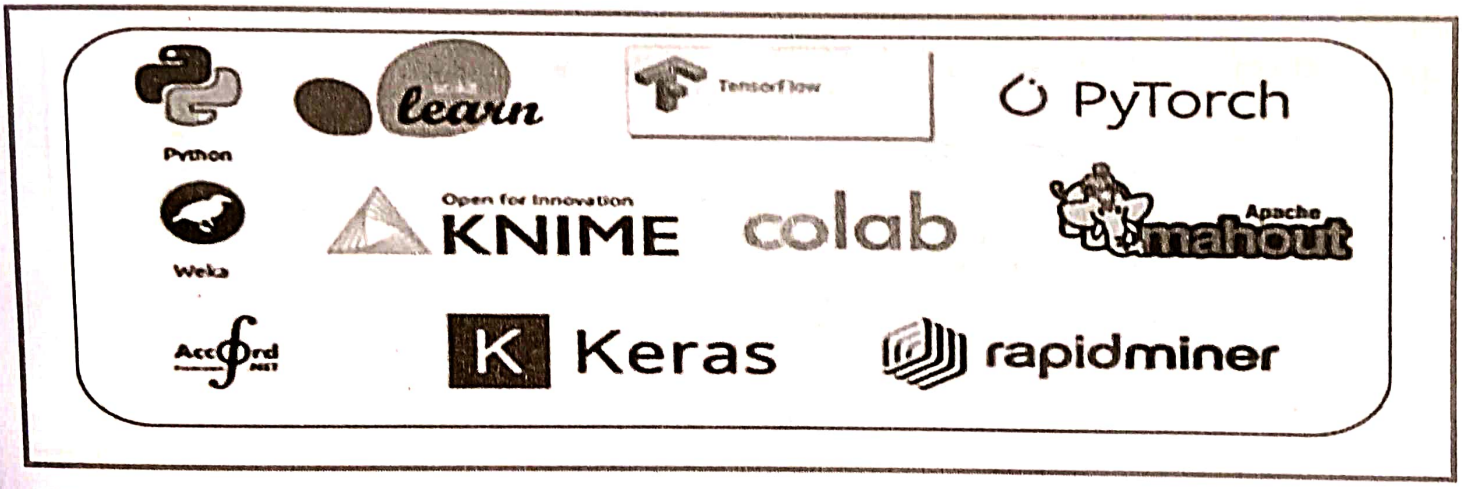
- **Data:** Machine learning relies on large datasets for model training and testing. The quality of the data significantly impacts the accuracy of the results. In many cases, data is continuously updated, requiring us to wait for new data to become available for further training and evaluation.
- **Time:** Machine learning algorithms need sufficient time to learn and evolve to achieve high accuracy and consistency. Additionally, operating these algorithms often demands considerable resources, including enhanced computational power, which may require upgrading existing systems.
- **Algorithm Selection:** Solving a machine learning problem often involves trying multiple algorithms to find the best fit. This process can be time-consuming and

requires manual testing of different models to identify the most accurate algorithm based on their performance.

- **Chances of High Error:** Machine learning algorithms can be chosen based on accuracy. We must apply the results to all algorithms. Data training and testing present significant challenges. Errors can be difficult to delete due to large amount of data. These errors can cause headaches for users.

2.1.8 MOST POPULAR MACHINE LEARNING SOFTWARE TOOLS

Machine learning software tools are essential for building, training, and deploying machine learning models. The choice of machine learning tool depends on the specific use case, data size, computational resources, and the familiarity with the tool. Python Libraries like **TensorFlow** and **PyTorch** are popular for deep learning, **Scikit-learn** is widely used for traditional machine learning, while cloud-based platforms like **Microsoft Azure** and **Google Cloud AI** offer end-to-end solutions for model building and deployment. Tools like **RapidMiner** and **KNIME** are great for users who prefer a visual interface to build machine learning workflows.



2.2 DEEP LEARNING

Deep learning, is a subset of machine learning, that aims to train computers to mimic human behavior through examples. It is a rapidly growing field that has revolutionized many industries, including healthcare, finance, and transportation. It has significantly advanced the field of AI, especially in areas such as image recognition, natural language processing (NLP), speech recognition, and autonomous driving.

Deep learning refers to the use of deep neural networks (DNNs), which are composed of several layers of artificial neurons. These networks aim to simulate the way the human brain processes information, using hierarchical layers to extract features and make predictions. The key difference between deep learning and traditional machine learning is the ability to learn from vast amounts of data without requiring manual feature extraction.

Deep learning models are particularly powerful because they can learn directly from raw data (such as images, audio, and text) without the need for preprocessing. As a result,

deep learning has revolutionized fields like computer vision, natural language processing, and speech recognition, where traditional algorithms struggled.

2.2.1 GENERIC MODEL OF DEEP LEARNING SYSTEMS

Deep learning systems typically follow a standard architecture consisting of multiple components that work together to process raw data and produce predictions or classifications. The core of the system is built around neural networks, and while the specific architecture may vary depending on the task (e.g., image classification, natural language processing), there are common components that define how deep learning systems operate. As shown in figure 2.5, the flow from raw data to deployment involves preprocessing, training, and fine-tuning to ensure that the model can generalize well to new data. The iterative process of training and optimization plays a crucial role in improving the model's accuracy and performance. Each component of the model is critical to ensuring the success of deep learning applications across industries like healthcare, finance, robotics, and more.

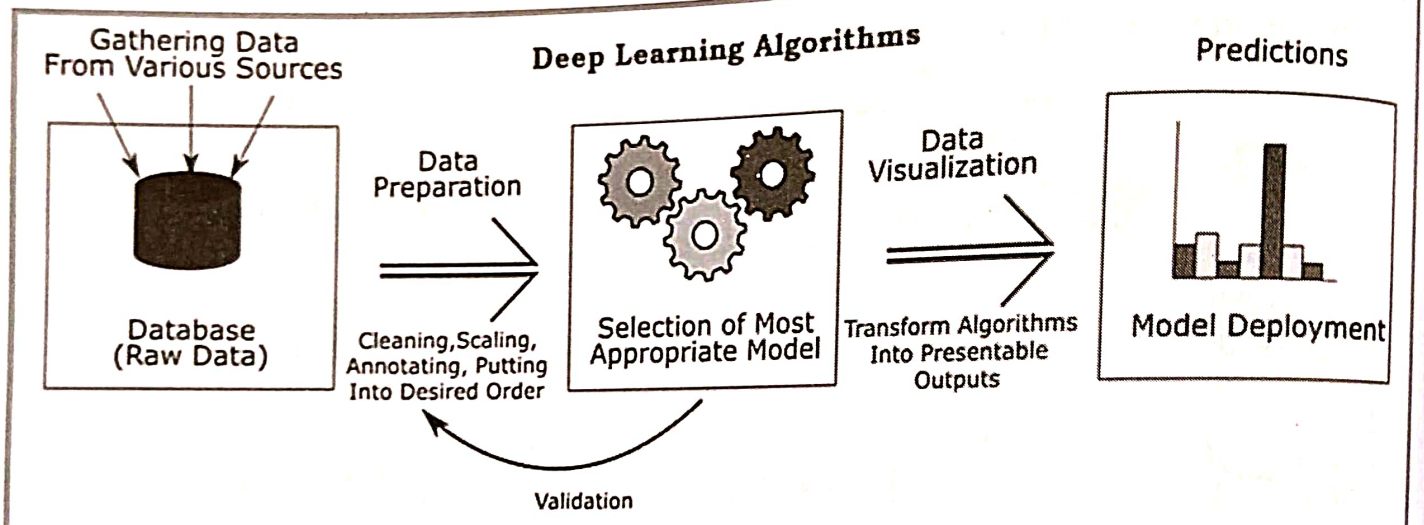


Fig. 2.5: Genetic Model of Deep Learning Systems

2.2.2 NEED OF DEEP LEARNING

Deep learning is needed because it addresses limitations in traditional machine learning and rule-based systems, offering the ability to model complex patterns in data and achieve superior performance in tasks that were once considered challenging or infeasible. Its necessity arises from the increasing complexity of problems in various domains, the availability of large datasets, and advances in computational power. Some key points depicting the need of deep learning are :

1. **Handling Complex and High-Dimensional Data:** Deep learning is designed to process and understand complex, high-dimensional data. Traditional algorithms struggle with such data due to the "curse of dimensionality," which refers to the exponential increase in computational complexity as the dimensionality of the data grows.
2. **Automatic Feature Extraction:** Deep learning eliminates the need for manual feature engineering by automatically learning hierarchical representations of data. Unlike traditional machine learning methods, which require domain expertise to extract meaningful features, deep learning learns these features directly from raw data.

3. **Superior Performance in Key Domains:** Deep learning models have set new benchmarks in tasks like image recognition, speech recognition, natural language processing, and reinforcement learning. Their ability to outperform traditional approaches makes them indispensable for cutting-edge applications.
4. **Scalability with Big Data:** With the advent of big data, traditional methods struggle to leverage vast datasets effectively. Deep learning thrives in data-rich environments, as its performance improves with larger datasets and deeper architectures.
5. **Advances in Computational Power:** The development of high-performance GPUs, TPUs, and distributed computing frameworks has made training deep learning models feasible. These advances allow for the efficient training of complex models on massive datasets, enabling breakthroughs in AI.
6. **Success in Unstructured Data:** Deep learning excels in processing unstructured data, such as images, audio, text, and video, which make up a large proportion of the data generated today. Traditional algorithms are limited in their ability to handle such data effectively.
7. **End-to-End Learning:** Deep learning allows for end-to-end learning, where the model learns to map raw input data directly to output predictions without needing intermediate steps or handcrafted rules.
8. **Adaptability to Diverse Applications:** Deep learning models are highly adaptable and can be fine-tuned for various applications by retraining them on new datasets. This flexibility makes them suitable for a wide range of industries.
9. **Handling Sequential and Temporal Data:** Deep learning architectures like Recurrent Neural Networks (RNNs) and Transformers are specifically designed for sequential and temporal data, such as time series, speech, and language. These architectures can model dependencies and relationships over time, which traditional methods struggle to capture.
10. **Generative Capabilities:** Deep learning enables generative tasks, where models can create new data similar to the training data. This capability is essential for creative applications, data augmentation, and simulations.

2.2.3 RELATION WITH MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Deep learning, machine learning (ML), and artificial intelligence (AI) are interconnected fields in the hierarchy of intelligent systems. **Artificial Intelligence** is the broadest domain encompassing technologies that enable machines to mimic human-like intelligence, such as decision-making, problem-solving, and learning. **Machine Learning**, a subset of AI, focuses on algorithms that allow machines to learn patterns from data and improve their performance without explicit programming. Within ML, **Deep Learning** is a specialized subfield that uses artificial neural networks inspired by the human brain to solve complex problems.

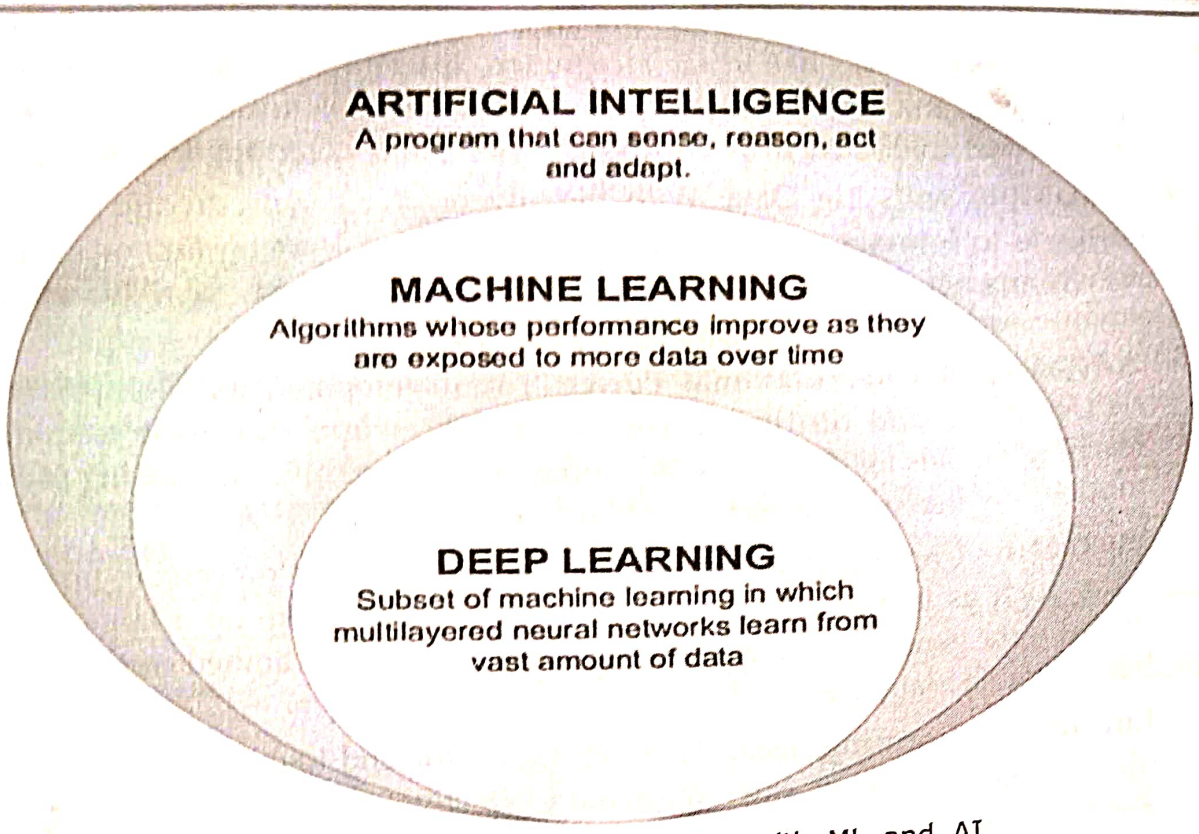


Fig. 2.6: Relation of Deep Learning with ML and AI.

While machine learning typically relies on feature engineering and simpler algorithms (e.g., decision trees, support vector machines), deep learning automates feature extraction by learning hierarchical representations directly from raw data. This makes deep learning particularly effective for tasks involving high-dimensional and unstructured data, such as images, speech, and text.

The relationship is symbiotic: AI is the overarching goal, ML provides the tools to achieve that goal, and deep learning offers cutting-edge techniques to push the boundaries of AI capabilities. For instance, deep learning powers advancements in AI-driven technologies like autonomous vehicles, virtual assistants, and medical diagnostics, showcasing how these fields work together to build more sophisticated and intelligent systems.

2.2.4 SOME KEY TECHNIQUES IN DEEP LEARNING

There are several key techniques and architectures within deep learning, each suited to different types of tasks:

1. Artificial Neural Networks (ANNs)

Artificial neural networks are the foundation of deep learning. They consist of layers of nodes (neurons), with each node representing a computational unit. The architecture typically includes:

- **Input Layer:** The initial layer where data enters the network.
- **Hidden Layers:** Intermediate layers where computation occurs. These layers learn features from the input data.

- **Output Layer:** The final layer, where the prediction or classification result is produced.

2. Convolutional Neural Networks (CNNs)

CNNs are designed for processing grid-like data, such as images. They are based on the concept of convolution, where filters (kernels) are applied to small patches of the input image. CNNs are particularly effective for tasks like image classification, object detection, and facial recognition.

Example: A CNN might be used to classify images of animals into categories like "cat," "dog," and "bird" by learning hierarchical features, such as edges, textures, and shapes.

Key Components of CNNs:

- **Convolutional Layer:** Performs convolutions with filters to detect features.
- **Pooling Layer:** Reduces the dimensionality of feature maps to reduce computational load.
- **Fully Connected Layer:** After feature extraction, fully connected layers make the final prediction.

3. Recurrent Neural Networks (RNNs)

RNNs are used for sequence data, where the output of the model depends on the previous elements in the sequence. This makes RNNs ideal for tasks involving time-series data, text, or speech. RNNs maintain an internal state to store information about previous inputs.

Example: In automatic speech recognition (ASR), an RNN might be used to process sequences of audio features over time and convert them into text.

2.2.5 ADVANTAGES OF DEEP LEARNING

- **Automatic Feature Extraction:** Deep learning eliminates the need for manual feature engineering by learning hierarchical representations directly from raw data. This reduces reliance on domain expertise and simplifies the development pipeline. Example: In image classification, a deep learning model can autonomously identify features like edges, textures, and objects without human intervention.
- **Handles Complex and Unstructured Data:** Deep learning excels at analyzing unstructured data such as images, videos, text, and audio. Traditional machine learning struggles with such data formats, whereas deep learning can process and understand them effectively. Example: Natural Language Processing (NLP) tasks like machine translation or speech recognition are powered by deep learning models.
- **Superior Accuracy and Performance:** Deep learning models often outperform traditional algorithms in tasks like image recognition, speech recognition, and natural language understanding due to their ability to model intricate patterns in

data. Example: Models like ResNet and BERT have achieved near-human or even superhuman performance in benchmarks.

- **End-to-End Learning:** Deep learning supports end-to-end learning, where raw inputs are directly mapped to outputs, eliminating the need for intermediate steps or separate components. Example: Autonomous driving systems use raw sensor data to directly predict steering commands.
- **Scales with Data:** Deep learning models improve as the size of the dataset increases, making them ideal for big data scenarios. Example: Platforms like Google and Facebook use deep learning to analyze massive user-generated datasets for recommendations and predictions.
- **Generative Capabilities:** Deep learning can generate new content, such as images, music, or text, through models like Generative Adversarial Networks (GANs) and transformers. Example: GANs create realistic synthetic images for design and animation industries.
- **Adaptability Across Domains:** Deep learning models can be fine-tuned for a variety of applications by retraining them on domain-specific datasets. Example: Pre-trained models like GPT can be adapted to tasks ranging from summarization to sentiment analysis.
- **Captures Non-Linear Relationships**: Deep learning models can learn highly complex, non-linear relationships between inputs and outputs that traditional models may fail to capture. Example: Predicting stock market trends based on historical data and external factors.
- **Continuous Improvements in Research:** The field of deep learning is rapidly advancing, with ongoing innovations in model architectures (e.g., Transformers, Vision Transformers) and optimization techniques, further enhancing its potential. Example: Models like ChatGPT represent leaps in conversational AI.
- **Supports Real-Time Applications:** Deep learning enables real-time decision-making and analysis, critical for applications like autonomous vehicles, fraud detection, and medical diagnostics. Example: Autonomous cars use deep learning for obstacle detection and collision avoidance in real-time.

2.2.6 CHALLENGES IN DEEP LEARNING

While deep learning has achieved remarkable successes, there are several challenges that researchers and practitioners face:

- **Data Requirements:** Deep learning models require large amounts of data for training, and obtaining labeled data can be expensive and time-consuming.
- **Interpretability:** Deep learning models, particularly deep neural networks, are often described as "black boxes" because their decision-making process is not easily interpretable. This lack of transparency can be a challenge in high-stakes fields like healthcare and finance.
- **Computational Resources:** Training deep learning models requires significant computational power, especially when dealing with large datasets. High-

performance GPUs and specialized hardware like TPUs (Tensor Processing Units) are often needed.

- **Overfitting:** Deep learning models are prone to overfitting, especially when trained on small datasets. Regularization techniques like dropout, data augmentation, and early stopping are used to mitigate this issue.
- **Time-consuming:** While working on sequential data depending on the computational resource it can take very large even in days or months.

2.2.7 APPLICATIONS OF DEEP LEARNING

Deep learning is being used in a wide variety of domains. Some of the most notable applications are as follows :

- **Computer Vision:** Computer vision tasks involve extracting meaningful information from images or videos. Deep learning has greatly advanced this field, with CNNs playing a central role. Examples- Image Classification, Object Detection, Face Recognition, etc.
- **Natural Language Processing (NLP) :** NLP involves the interaction between computers and human language. Deep learning techniques, especially transformers, have significantly improved tasks like text classification, translation, and summarization. Examples- Sentiment Analysis, Chatbots and Virtual Assistants, Text Generation, etc.
- **Speech Recognition:** Deep learning has enabled significant improvements in automatic speech recognition (ASR). RNNs, particularly LSTMs, and transformers are widely used in this area. Examples- Voice Assistants, Transcription, etc.
- **Autonomous Systems:** Autonomous vehicles rely heavily on deep learning techniques to perceive and understand their environment. Example- Self-Driving Cars: Using CNNs for image recognition and RNNs for decision-making, autonomous vehicles can navigate roads, recognize pedestrians, and make real-time decisions.
- **Healthcare:** Deep learning is transforming healthcare through applications in medical image analysis, drug discovery, and patient monitoring. Example- Medical Imaging (Detecting anomalies like tumors in X-rays or MRI scans using CNNs) and Predictive Analytics.

2.3 NATURAL LANGUAGE PROCESSING (NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that helps computers understand, interpret, and respond to human language in a way that feels natural. Simply put, NLP is what allows machines to make sense of the words we use every day—whether spoken or written—so they can interact with us more effectively.

Imagine talking to a virtual assistant like Siri, typing a query into Google, or using a chatbot to get customer support. All these systems rely on NLP to process what you say

or write, figure out what you mean, and provide a relevant response. This makes NLP a crucial technology in today's world, as it enables better communication between humans and machines.

At its core, NLP bridges the gap between human language, which is often ambiguous and filled with subtleties, and computer systems, which operate on strict logic and rules. By combining linguistic rules, statistical methods, and machine learning techniques, NLP systems can perform tasks like understanding the sentiment of a review, translating one language into another, summarizing large texts, and even generating coherent responses in a conversation.

The journey of NLP started decades ago with rule-based systems, which relied heavily on manually created grammar rules and dictionaries. While these systems worked for simple tasks, they struggled with the vast diversity and complexity of human language. As technology evolved, statistical methods and machine learning came into play, enabling computers to learn patterns in language from large datasets. Today, deep learning has revolutionized NLP, powering sophisticated systems like Google Translate, GPT-based chatbots, and personalized content recommendations.

2.3.1 NEED OF NLP

The need for NLP arises from the inherent complexity of human languages and the growing demand for intelligent systems to interact seamlessly with humans. This includes understanding text, speech, and context to automate tasks, extract insights, and facilitate communication. Need of NLP can be explained with following points :

- **Massive Growth of Unstructured Data:** A significant portion of data generated today is unstructured, primarily in the form of text and speech. Social media posts, emails, customer reviews, blogs, and online documents are examples of such data. NLP helps in processing and deriving insights from this unstructured information. Example: Sentiment analysis of customer reviews to understand opinions about a product.
- **Bridging the Human-Computer Interaction Gap:** Traditional systems require structured inputs like forms or programming commands. NLP enables more intuitive interactions using natural language, making systems accessible to non-technical users. Example: Virtual assistants like Siri, Alexa, and Google Assistant use NLP to understand voice commands.
- **Automation of Repetitive Language-Intensive Tasks:** Many tasks, such as document summarization, email classification, and language translation, can be automated using NLP, saving time and resources. Example: Automatic spam detection in emails.
- **Handling Multilingual Needs:** With globalization, businesses and platforms need to operate across linguistic barriers. NLP-powered translation systems help bridge language gaps. Example: Google Translate facilitates communication across multiple languages.

- **Decision Making and Insight Generation:** NLP is crucial for analyzing textual data to support decision-making in business, healthcare, finance, and more. Example: Extracting key insights from research papers in drug discovery.
- **Speech-to-Text and Text-to-Speech Applications:** NLP powers speech recognition systems, making communication and accessibility easier for users who prefer voice-based interactions. Example: Automatic transcription services like Otter.ai.
- **Understanding Context and Intent:** Human language is rich in context, ambiguity, and idiomatic expressions. NLP helps systems infer intent and meaning accurately, which is essential for conversational AI. Example: Chatbots understanding whether a query is a complaint or a request for information.

2.3.2 NATURAL LANGUAGE PROCESSING TECHNIQUES

The two major techniques used in natural language processing include:

2.3.2.1. Syntax Analysis

Syntax refers to the structure of words in a sentence to ensure that the grammar is correct. NLP analyzes a language's meaning through syntax and grammatical rules. These steps are foundational for many Natural Language Processing tasks, enabling computers to process and understand human language effectively.

Various syntax NLP techniques are:

- **Lemmatization:** Lemmatization is the process of converting different inflected forms of a word into its base or dictionary form, called a lemma. This helps standardize the text for easier analysis. Example: The words "running", "ran", and "runs" are all reduced to the lemma "run".
- **Word Segmentation:** Word segmentation involves splitting a continuous stream of text into distinct words or units, especially in languages where words are not clearly separated by spaces. Example: The text "I like learning" is segmented into "I/like/learning".
- **Morphological Segmentation:** Morphological segmentation breaks words into smaller units called morphemes, which are the smallest meaningful components of a word. Example: The word "unbelievable" is segmented into morphemes: "un-" (prefix meaning "not"), "believe" (root word), and "-able" (suffix meaning "capable of").
- **Part-of-Speech (POS) Tagging:** POS tagging assigns a part of speech (e.g., noun, verb, adjective) to each word in a text, helping in understanding its grammatical role.

  Example: In the sentence "She runs fast":

  "She" ⟶ Pronoun (PRON)

  "runs" ⟶ Verb (VERB)

  "fast" ⟶ Adverb (ADV)
- **Sentence Breaking:** Sentence breaking, also known as sentence boundary detection, identifies the boundaries between sentences in a large piece of text.

Example: In the text "I love NLP. It is fascinating!", sentence breaking would split it into two sentences:

"I love NLP."

"It is fascinating!"

- **Parsing:** Parsing involves analyzing the grammatical structure of a sentence, typically resulting in a parse tree or syntactic structure. Example: For the sentence "The cat sat on the mat", parsing might produce a tree structure showing:

  Subject: "The cat"

  Verb: "sat"

  Prepositional phrase: "on the mat"

- **Stemming:** Stemming reduces words to their root form by removing suffixes or prefixes, often without considering the actual meaning. Example: The words "playing", "played", and "plays" are reduced to the stem "play".

2.3.2.2 Semantic Analysis

Semantic deals with the use and the meaning of various words. These techniques enable computers to understand and generate human-like language, improving their ability to process and interact with semantic content in a meaningful way.

- **Named Entity Recognition (NER) :** NER is the process of identifying specific entities within a text and classifying them into predefined categories, such as names of people, locations, organizations, dates, etc. Example: In the sentence "Barack Obama was born in Hawaii on August 4, 1961.", NER would identify:

  "Barack Obama" as a Person (PERSON)

  "Hawaii" as a Location (LOCATION)

  "August 4, 1961" as a Date (DATE)

- **Word Sense Disambiguation (WSD) :** WSD is the process of determining the meaning of a word based on its context in a sentence. This helps resolve ambiguities when a word has multiple meanings. Example: The word "bank" has different meanings depending on the context:

  In "I went to the bank to deposit money," it refers to a financial institution.

  In "The boat was near the river bank," it refers to the land alongside a river.

- **Natural Language Generation (NLG) :** NLG involves converting structured data or semantic intentions into human-readable text. It typically involves transforming data from a database into natural language that can be understood by humans. Example: Given structured data such as "Temperature: 72°F, Weather: Clear, Wind Speed: 5 mph", NLG might generate the sentence:

  "The weather is clear with a temperature of 72°F and wind speeds of 5 mph."

2.3.3 APPLICATIONS OF NLP

Natural Language Processing (NLP) has numerous applications across various industries. These applications leverage NLP techniques to make interactions with computers

ıore intuitive and to automate complex language tasks. Below are some of the key pplications of NLP:

- **Machine Translation:** NLP is used to translate text or speech from one language to another.

  Examples:

  - **Google Translate:** Translates text and web pages between multiple languages.
  - **DeepL:** Offers high-quality translations by leveraging deep learning techniques for better accuracy.

- **Sentiment Analysis:** NLP techniques are used to detect and categorize emotions expressed in text, such as positive, negative, or neutral sentiments.

  Examples:

  - **Social Media Monitoring:** Analyzing tweets or posts on platforms like Twitter and Facebook to gauge public sentiment about a product or event.
  - **Customer Feedback Analysis:** Evaluating customer reviews or feedback to determine the overall sentiment towards a brand or service.

- **Chatbots and Virtual Assistants:** NLP powers chatbots and virtual assistants by enabling them to understand and respond to user queries in natural language.

  Examples:

  - **Siri:** Apple's voice assistant that interprets spoken commands and questions.
  - **Amazon Alexa:** A voice-controlled virtual assistant that can perform tasks like playing music, setting reminders, and controlling smart home devices.

- **Information Retrieval:** NLP helps in retrieving relevant information from large datasets or documents, often through search engines or recommendation systems.

  Examples:

  - **Search Engines:** Google uses NLP to process search queries and return relevant web pages.
  - **Enterprise Search Systems:** Systems that search and retrieve documents from an organization's internal database.

- **Text Summarization:** NLP algorithms condense long pieces of text into shorter summaries while preserving essential information.

  Examples:

  - **News Aggregators:** Automatically generating summaries of news articles for quick reading.
  - **Research Paper Summaries:** Summarizing lengthy academic papers or reports for quick comprehension.

- **Named Entity Recognition (NER) :** NER is used to identify and classify entities in text, such as names of people, organizations, dates, and locations.

Examples:

- **Document Processing:** Extracting key information from contracts, legal documents, or business reports.
- **Social Media Monitoring:** Identifying mentions of celebrities, companies, or locations in tweets and posts.

- **Speech Recognition:** NLP is used to convert spoken language into text, enabling voice-based interaction with computers.

  Examples:

  - **Speech-to-Text Applications:** Dictation software like Dragon Naturally Speaking converts spoken words into written text.
  - **Voice Search:** Google's voice search interprets spoken queries and returns results without typing.

- **Question Answering (QA) Systems:** NLP helps machines understand and respond to natural language questions, often drawing from a knowledge base or external sources.

  Examples:

  - **Siri or Google Assistant:** Answering questions based on knowledge repositories and real-time data.
  - **IBM Watson:** A powerful AI system that can answer complex questions in fields like healthcare and business.

- **Language Generation:** NLP techniques enable the generation of human-like text based on certain inputs, such as data, text prompts, or context.

  Examples:

  - **Content Creation:** Automatically generating reports, blogs, and articles.
  - **Creative Writing:** Tools like GPT-4 can generate creative text such as poems, stories, and dialogue.

- **Text-to-Speech (TTS) and Speech-to-Text (STT) :** NLP systems can convert written text into spoken words and vice versa, enabling a more interactive experience for users.

  Examples:

  - Screen Readers, Google Assistant and Amazon Alexa
  - Transcription Services, Voice Commands

- **Healthcare Applications:** NLP is used to extract valuable information from medical texts, such as patient records, research papers, and clinical notes.

  Examples:

  - **Medical Text Mining:** Extracting conditions, symptoms, and treatment data from unstructured clinical notes.
  - **Clinical Decision Support:** Using NLP to identify relevant medical literature for doctors and healthcare professionals.

2.3.4 ADVANTAGES OF NLP

Natural Language Processing (NLP) offers several advantages that can significantly enhance various processes across different industries. Here are some of the key benefits:

- Enhances the accuracy and efficiency of document processing.
- Allows organizations to implement chatbots for enhanced customer support.
- Enables automatic generation of concise summaries from lengthy, complex texts.
- Facilitates the analysis of both structured and unstructured data.
- Allows voice-activated personal assistants, like Alexa, to comprehend spoken language.
- Simplifies sentiment analysis for organizations.
- Helps organizations gain deeper insights into lead generation, social media posts, surveys, and reviews.
- Provides advanced analytics that offer valuable insights previously inaccessible due to the large volume of data.

2.3.5 CHALLENGES OF NLP

Natural Language Processing (NLP) faces numerous challenges, most of which arise from the inherent complexity and constantly evolving nature of human language. Some of the key challenges include:

- **Voice Tone and Inflection:** One of the difficulties in NLP is the challenge of capturing tone and inflection in speech. For example, understanding sarcasm, which relies on tone and context, can be particularly tricky for NLP systems. Additionally, the meaning of a sentence can shift depending on which word or syllable is emphasized. NLP algorithms may miss these subtle changes, which can be crucial for accurate interpretation. Variations in accent can also present problems, as algorithms may struggle to recognize speech patterns across different linguistic regions.
- **Evolving Language:** Language is continuously evolving, and the way people use it changes over time. While language follows certain rules, these rules are not static; they shift with cultural trends, societal changes, and technological developments. As a result, computational rules that are effective today may become outdated, and NLP systems must constantly adapt to new language patterns and usages.
- **Bias:** NLP systems can unintentionally inherit biases from their training data. If the data used to train an NLP model contains biased information, the model may reproduce these biases in its outputs. This is particularly concerning in areas such as healthcare and hiring, where biases could result in discrimination or unequal treatment of individuals based on gender, ethnicity, or other factors.
- **Precision:** Unlike computers, which are designed to process information in highly structured, unambiguous programming languages or limited, clearly articulated voice commands, human language is often imprecise and ambiguous. The structure

of human speech can vary based on many factors, such as slang, regional dialects, and social context, making it difficult for machines to fully understand and process spoken language accurately.

- **Understanding Context:** Context is crucial for proper language understanding, as meaning can change based on the situation in which it is used. However, NLP models often struggle to capture the full context of a conversation or text.
- **Computational Complexity:** Processing large amounts of text data or complex linguistic structures can be computationally expensive and time-consuming. Deep learning models used in NLP can require significant computational resources, especially when dealing with large datasets or real-time applications.

2.4 COMPUTER VISION

Computer vision is defined as a field of AI that enables computers and systems to derive meaningful information from digital images, videos, and other visual inputs. These systems process visual data to make sense of the environment and take actions based on their interpretations, much like the human visual system.

In essence, computer vision mimics how humans use their eyes and brain to recognize objects, interpret scenes, and make decisions based on visual data. It involves the development of algorithms that can analyze and understand the content of images or videos to detect patterns, objects, and other features.

2.4.1 HOW DOES COMPUTER VISION WORK?

Computer vision is all about pattern recognition, which includes training machine systems to understand visual data like images and videos. To begin, machines are given a large amount of visually labeled data to train with. This labeled data allows the machine to identify different patterns in all of the data points and relate them to those labels. Assume we provide visual data containing millions of dog images. In that case, the computer learns from the data by analyzing each photo, shape, distance between shapes, color, and so on, identifying patterns that are similar to dogs and generating a model. As a result, this computer vision model can now correctly detect whether the image contains a dog or not for each.

2.4.2 NEED OF COMPUTER VISION

The need for computer vision arises from its potential to transform various industries by automating processes that traditionally relied on human vision and judgment. Some of the key reasons for its importance include:

- **Automation**: In many fields, such as manufacturing and logistics, computer vision can replace human workers in repetitive tasks that require visual assessment, reducing labor costs and errors.
- **Efficiency**: By automating the process of visual inspection, computer vision systems can process large volumes of data much faster than humans, improving efficiency in applications like quality control, inventory management, and security.

- **Improved Accuracy:** Machines do not suffer from human limitations such as fatigue or distraction, meaning they can perform tasks that require constant attention with greater consistency and accuracy.
- **Cost Reduction:** In areas like surveillance and healthcare, where human involvement can be expensive or inefficient, computer vision can help reduce operational costs.
- **Enhanced Safety:** In autonomous vehicles or drones, computer vision is crucial for navigating environments safely, detecting obstacles, and making real-time decisions.

2.4.3 TYPES OF COMPUTER VISION TASKS

Computer vision can be categorized into several types of tasks, each focused on a specific area of visual understanding. These include:

(a) **Image Classification:** Image classification involves categorizing an image into a predefined class or category. For example, a computer vision system might classify an image of a cat as belonging to the "cat" category.

(b) **Object Detection:** Object detection goes a step further by not only classifying objects but also locating them within an image. This involves drawing bounding boxes around detected objects, as seen in systems that detect faces or vehicles.

(c) **Semantic Segmentation:** Semantic segmentation is the task of assigning a class label to each pixel in an image. It helps to understand the context and relationships between different regions in an image.

(d) **Instance Segmentation:** Instance segmentation is a more advanced version of segmentation, where the system not only labels each pixel but also differentiates between different instances of the same object category (e.g., differentiating between two cars in an image).

(e) **Facial Recognition:** Facial recognition identifies and verifies individuals by analyzing facial features. It is widely used in security systems, social media platforms, and personal devices.

(f) **Optical Character Recognition (OCR):** OCR is used to convert images of text into machine-readable text. This technology is applied in document scanning and automated data entry.

(g) **Motion Analysis:** Motion analysis involves tracking the movement of objects or individuals in video feeds. It's used in applications such as surveillance and autonomous vehicles.

2.4.4 APPLICATIONS OF COMPUTER VISION

Computer vision has diverse applications across industries, where it adds value by automating tasks and improving decision-making. Key applications include:

- **Autonomous Vehicles :** In self-driving cars, computer vision is used for object detection, lane recognition, and navigation. The system needs to interpret surroundings in real-time, such as identifying pedestrians, traffic signs, and other vehicles.

- **Healthcare and Medical Imaging:** Computer vision is applied in medical imaging to assist in diagnosing diseases. For example, algorithms can be used to detect tumors in radiology scans, classify cells in pathology slides, or monitor patient vitals.
- **Security and Surveillance:** In security, computer vision is used for surveillance, recognizing faces in public places or detecting suspicious behavior in real-time video feeds.
- **Retail and E-commerce:** Retailers use computer vision for inventory management, customer behavior analysis, and even cashier-less checkout systems (like Amazon Go), which rely on cameras and computer vision to track products customers take.
- **Manufacturing and Quality Control:** In manufacturing, computer vision systems are used for automated inspection and quality control, identifying defective products on production lines.
- **Augmented Reality (AR) :** AR applications use computer vision to track real-world objects and overlay digital information on top of them, such as in gaming, education, or marketing.
- **Agriculture:** In agriculture, computer vision is used for crop monitoring, pest detection, and automated harvesting by analyzing images captured by drones or other sensors.
- **3D model building:** 3D model building or 3D modeling is a technique to generate a 3D digital representation of any object or surface using the software. In this field also, computer vision plays its role in constructing 3D computer models from existing objects.

2.4.5 ADVANTAGES OF COMPUTER VISION

The primary advantages of computer vision are as follows:

- **Automation of Repetitive Tasks**: CV systems can handle repetitive and time-consuming tasks with greater accuracy and efficiency than humans.
- **Enhanced Accuracy and Precision**: Computer vision systems can detect details that human eyes may miss, leading to more accurate decision-making.
- **Scalability**: Computer vision applications can be scaled to handle large amounts of visual data, such as processing thousands of images or videos per day.
- **Real-time Processing**: With the help of advanced algorithms, computer vision can be used for real-time analysis and decision-making.
- **Cost Savings**: Automation reduces labor costs and improves operational efficiency, leading to cost savings in various industries.

2.4.6 CHALLENGES OF COMPUTER VISION

Despite its many advantages, computer vision faces several challenges:

- **Complexity**: Developing accurate computer vision models requires expertise in machine learning, image processing, and domain-specific knowledge.

- **Data Dependency**: The performance of computer vision systems relies heavily on the quality and diversity of the data used for training.
- **High Computational Resources**: Computer vision tasks, particularly deep learning, can be computationally expensive, requiring high-end hardware like GPUs.
- **Bias and Fairness**: Computer vision models can inherit biases from training data, leading to biased outcomes, especially in sensitive applications like facial recognition.
- **Ambiguity in Visual Data**: Objects in images can be occluded, distorted, or partially visible, making it difficult for computer vision systems to detect or recognize them accurately.

2.5 ROBOTICS

Robotics is a multidisciplinary field of engineering and science that focuses on the design, construction, operation, and application of robots. A robot is a programmable machine capable of carrying out a complex series of actions automatically, often with the goal of performing tasks that are dangerous, repetitive, or otherwise difficult for humans to carry out. Robotics blends elements from fields such as mechanical engineering, electrical engineering, computer science, artificial intelligence (AI), and control engineering to create systems that can interact with their environment and make decisions based on sensor input.

The development of robotics aims to improve the efficiency, safety, and versatility of machines in a wide range of applications. Over the years, robotics has evolved from simple, task-specific machines to highly advanced systems capable of performing intricate and adaptive tasks.

Robotics is defined as the branch of technology that deals with the design, construction, operation, and use of robots. Robots are autonomous or semi-autonomous machines that can perform a variety of tasks through a set of pre-programmed commands or real-time sensor feedback.

2.5.1 KEY COMPONENTS OF A ROBOT

A robot typically consists of several key components that allow it to interact with its environment and perform tasks:

- **Actuators**: Actuators are the motors or engines that allow a robot to move or perform actions. They could be electric motors, hydraulic actuators, or pneumatic actuators.
- **Sensors**: Sensors provide input to the robot, enabling it to perceive its environment. These can include cameras, proximity sensors, gyroscopes, accelerometers, temperature sensors, and force sensors.
- **Control System**: The control system is the brain of the robot, responsible for interpreting sensor data and sending commands to the actuators. It can be a

simple microcontroller or a more sophisticated computing platform with AI capabilities.

- **Power Supply**: Robots require a power source to operate. This could be a battery, fuel cells, or an external power supply.
- **End Effector**: The end effector is the tool or appendage that interacts with the robot's environment. This could be a gripper, a welding tool, a camera, or a robotic arm, depending on the robot's intended task.
- **Software**: The software controls the robot's movements, decision-making, and response to environmental changes. It is often based on algorithms related to robotics and artificial intelligence, such as pathfinding, object recognition, and motion planning.

2.5.2 TYPES OF ROBOTS

Robots can be classified into different types based on their functionality, mobility, and design. Some of the most common types include:

(a) **Industrial Robots:** Industrial robots are primarily used in manufacturing and production environments. These robots are typically fixed in place and used for tasks such as welding, painting, assembly, packaging, and material handling. They are often characterized by their precision, speed, and ability to perform repetitive tasks.

Example: The robotic arms used in automotive assembly lines to weld car parts together.

(b) **Service Robots:** Service robots are designed to assist humans in tasks outside of industrial settings. These robots can be used in environments like healthcare, hospitality, and personal assistance.

Example: Robotic vacuum cleaners like the Roomba or robots used in hospitals for patient delivery.

(c) **Autonomous Mobile Robots (AMRs):** AMRs are robots that can navigate and operate autonomously in dynamic environments, usually equipped with a range of sensors, cameras, and AI systems to map their surroundings and perform tasks.

Example: Autonomous delivery robots that transport goods in warehouses or self-driving cars.

(d) **Humanoid Robots:** Humanoid robots are designed to resemble the human body, with a torso, limbs, and often a head. These robots aim to replicate human behavior and movements and are often used in research, entertainment, and personal assistance.

Example: Robots like Honda's ASIMO or Softbank's Pepper, which can interact with humans.

(e) **Educational Robots:** Educational robots are designed to teach concepts related to robotics, programming, and engineering. They are commonly used in schools and universities as learning tools.

Example: LEGO Mindstorms, a popular robotics kit used for teaching children how to build and program robots.

2.5.3 TECHNIQUES AND TECHNOLOGIES IN ROBOTICS

The field of robotics integrates various technologies and techniques, including:

(a) **Artificial Intelligence and Machine Learning:** AI and machine learning algorithms enable robots to learn from data and adapt to new situations. This can include learning to recognize objects, adapt to unforeseen circumstances, and optimize their movements over time.

Example: A robot using machine learning to improve its object recognition capabilities based on a growing database of images.

(b) **Computer Vision:** Computer vision involves the use of cameras and sensors to provide visual feedback to robots. This technology helps robots "see" their environment, detect obstacles, recognize objects, and make decisions based on visual data.

Example: Robots that use computer vision to avoid obstacles while navigating a room or warehouse.

(c) **Path Planning and Motion Control:** Path planning algorithms allow robots to navigate complex environments efficiently. Robots use these algorithms to determine the best path from one point to another while avoiding obstacles and considering factors like energy consumption and speed.

Example: A robot used in autonomous vehicles plans a route while avoiding traffic and pedestrians.

(d) **Robotics Kinematics:** Robotics kinematics refers to the study of motion and the movement of robot parts. It involves understanding the mathematical relationships between different parts of the robot and ensuring they move in a coordinated manner.

Example: A robotic arm moving smoothly to pick up an object and place it in a specific location.

2.5.4 APPLICATIONS OF ROBOTICS

Robotics has a wide range of applications across industries and fields, some of which include:

(a) **Manufacturing and Automation:** Robots are widely used in manufacturing to automate processes such as assembly, painting, welding, and material handling. Robots improve efficiency, accuracy, and safety by performing repetitive tasks.

Example: Robots in car manufacturing that assemble car parts on a production line.

(b) **Healthcare and Medical Robotics:** Medical robots are used for surgery, rehabilitation, and patient care. These robots are designed to enhance the precision of medical procedures, assist doctors during surgery, and provide therapy to patients.

Example: Surgical robots like the Da Vinci Surgical System, which assist surgeons in performing minimally invasive surgeries.

(c) **Agriculture:** Robots in agriculture are used to automate tasks like planting, harvesting, and crop monitoring. These robots help increase productivity and reduce the reliance on manual labor.

Example: Drones and autonomous vehicles used for monitoring crops and applying pesticides.

(d) **Space Exploration:** Robots play a crucial role in space exploration by assisting in tasks that are too dangerous or impractical for humans, such as operating on the surface of other planets or repairing satellites.

Example: The Mars rovers, which explore the Martian surface and send data back to Earth.

(e) **Service Industry:** Robots are used in the service industry for tasks like customer interaction, cleaning, and delivery. Service robots improve the efficiency and customer experience in various industries, including hospitality and retail.

Example: Robots that deliver food in restaurants or clean hotel rooms.

(f) **Military and Defence:** Robots are used in military applications for bomb disposal, reconnaissance, and search-and-rescue operations. They enhance safety by performing dangerous tasks remotely.

Example: Bomb disposal robots that handle hazardous materials or defuse explosive devices.

2.5.5 ADVANTAGES OF ROBOTICS

- **Increased Efficiency:** Robots can perform tasks faster and more accurately humans, especially in repetitive and precision-demanding tasks.
- **Safety:** Robots can perform dangerous jobs in hazardous environments, reducing the risk of injury to human workers.
- **Cost Reduction:** Automation using robots can reduce labor costs and improve overall operational efficiency.
- **Precision and Accuracy:** Robots are highly accurate, performing tasks with consistent precision, which is essential in fields like surgery and manufacturing.
- **Availability:** Robots can work 24/7 without the need for breaks, increasing productivity and reducing downtime.

2.5.6 DISADVANTAGES AND CHALLENGES OF ROBOTICS

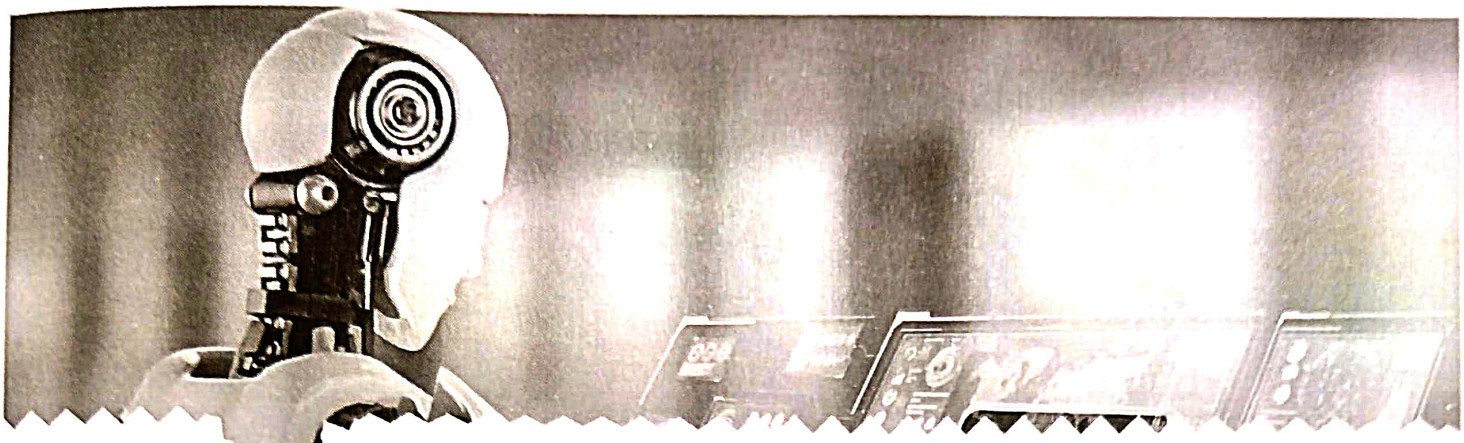
- **High Initial Cost:** The design, development, and implementation of robots can be costly, requiring significant investment.
- **Job Displacement:** The widespread use of robots in industries may lead to job loss, particularly for workers performing routine or manual tasks.
- **Complexity:** Designing robots capable of performing intricate tasks can be technologically complex and time-consuming.

- **Limited Flexibility:** Many robots are specialized for specific tasks, and adapting them for new functions can be challenging.

QUESTIONS

1. Explain the major technologies of AI ?
2. How Machine learning, Deep learning and NLP related to AI ?
3. How machine learning is beneficial for us ? Give its various applications.
4. Differentiate between supervised, unsupervised and reinforcement learning.
5. What is the need of deep learning ? Discuss its key techniques.
6. Explain the techniques used in NLP giving some real world applications.
7. What is computer vision ? Give its applications.

☆☆☆☆☆



CHAPTER

3 PROBLEM SOLVING AND APPLICATIONS OF AI

CHAPTER AT A GLANCE

3.1 Applications of AI	3.5 Problem Solving
3.2 Applications of AI in Healthcare	3.6 Agents and Environment
3.3 Applications of AI in Education	3.7 Ethical Issues and Future of AI
3.4 Application of AI in Agriculture	

3.1 APPLICATIONS OF AI

AI is transforming industries and revolutionizing how we interact with technology. Its ability to analyze vast amounts of data, learn from patterns, and making autonomous decisions makes it a powerful tool across various domains. It can effectively address complicated problems in a variety of areas, including healthcare, entertainment, banking, education, etc., resulting it to become indispensable in the modern era. AI is improving the speed and comfort of our daily lives.

1. Finance:

AI is transforming the finance industry by enhancing security, improving customer service, and optimizing financial operations. In finance, AI can be applied for the following tasks:

- **Fraud Detection:** AI algorithms analyze transaction patterns to identify and prevent real-time fraudulent activities.
- **Algorithmic Trading:** AI-driven trading systems use historical data and market trends to execute high-frequency trades, optimizing investment strategies.

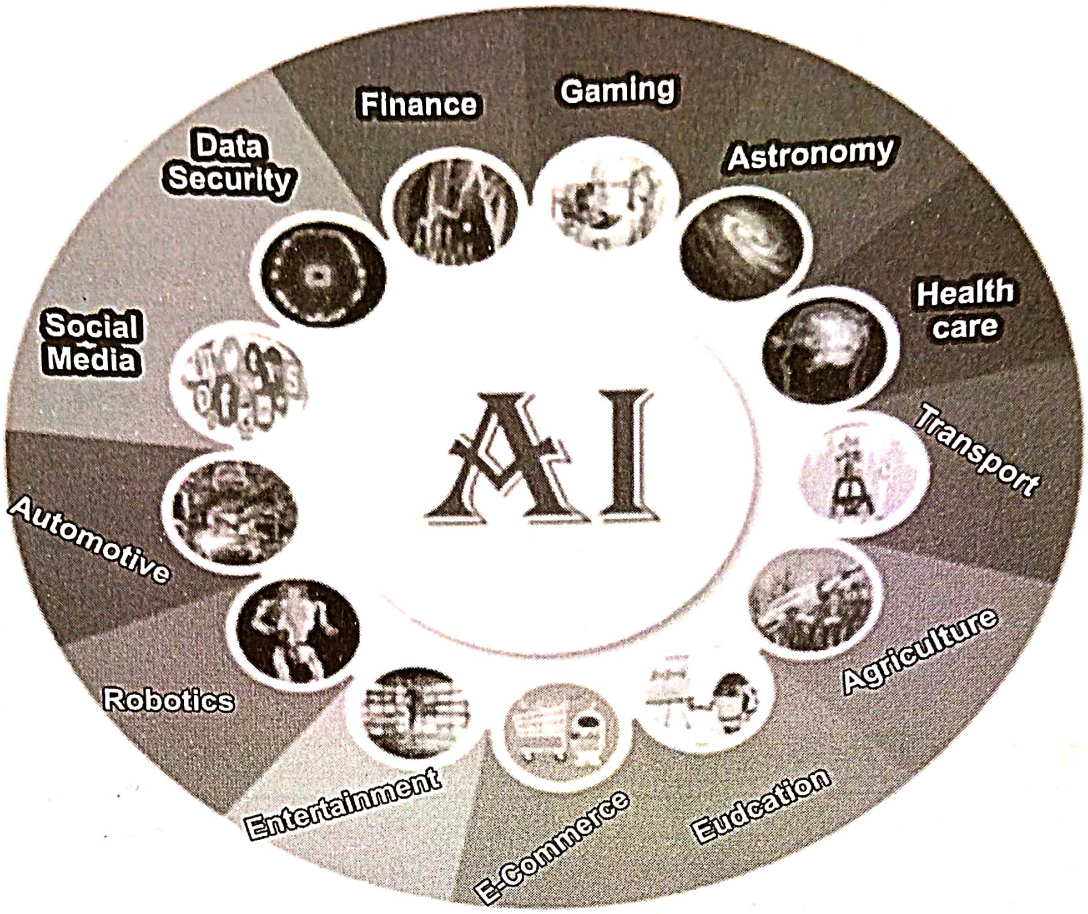


Fig. 3.1: Application Areas of A2.

- **Customer Service:** AI chatbots and virtual assistants handle customer inquiries, process transactions, and provide financial advice, improving efficiency and customer satisfaction.
- **Credit Scoring:** Machine learning models assess credit worthiness by analyzing various data points, offering more accurate and fair credit scores.

2. Retail

AI is reshaping the retail industry by enhancing customer experiences, optimizing inventory management, and driving sales. AI can be applied for the following tasks:

- **Personalized Recommendations:** AI algorithms analyze customer behavior and preferences to provide personalized product recommendations, increasing sales and customer loyalty.
- **Inventory Management:** AI systems predict demand and optimize inventory levels, reducing waste and ensuring products are available when needed.
- **Chatbots and Virtual Assistants:** AI-powered chatbots assist customers with product inquiries, order tracking, and returns, providing efficient and personalized customer service.
- **Visual Search:** AI enables customers to search for products using images, making finding items that match their preferences easier.

3. Transportation

AI in automotive industry is revolutionizing transportation by improving safety, efficiency, and convenience. AI has been widely applied for following tasks:

- **Autonomous Vehicles:** AI-powered self-driving cars and trucks use sensors and machine learning to navigate and make driving decisions, reducing accidents and improving traffic flow.
- **Predictive Maintenance:** AI analyzes data from vehicles and infrastructure to predict maintenance needs, preventing breakdowns and reducing downtime.
- **Route Optimization:** AI algorithms optimize routes for logistics companies, reducing fuel consumption and delivery times.
- **Traffic Management:** AI systems monitor and manage traffic flow in real-time, reducing congestion and improving overall transportation efficiency.

4. Manufacturing

AI in manufacturing has been enhancing production processes, quality control, and supply chain management. Manufacturing field use AI in:

- **Predictive Maintenance:** AI predicts equipment failures before they occur, allowing for proactive maintenance and reducing downtime.
- **Quality Control:** Machine learning models analyze production data to detect defects and ensure product quality.
- **Robotics:** AI-powered robots precisely perform repetitive and dangerous tasks, increasing productivity and worker safety.
- **Supply Chain Optimization:** AI optimizes supply chain operations by predicting demand, managing inventory, and improving logistics.

5. Entertainment

AI is reshaping the entertainment industry by creating new content, enhancing user experiences, and optimizing production processes. Major use of AI is done for the following tasks:

- **Content Recommendation:** AI algorithms analyze user preferences and behavior to provide personalized content recommendations on streaming platforms.
- **Game Development:** AI generates realistic characters and environments, enhancing the gaming experience.
- **Video Editing:** AI automates video editing processes, such as cutting, filtering, and adding effects, speeding up production times.
- **Music Composition:** AI composes music by analyzing existing compositions and creating new pieces in various styles and genres.

6. Energy

AI optimizes energy production, distribution, and consumption, promoting sustainability.

- **Smart Grids:** AI balances energy supply and demand in real time to prevent outages.
- **Energy Efficiency:** Tools like Google's DeepMind reduce energy consumption in data centers.
- **Predictive Maintenance:** AI monitors equipment in power plants to detect faults early.
- **Energy Trading:** AI algorithms automate energy trading in dynamic markets.

7. Law and Legal Services

AI streamlines legal processes, making services more accessible and efficient.

- **Contract Analysis:** Tools like Kira Systems review contracts and identify key clauses.
- **Legal Research:** Platforms like ROSS Intelligence use AI to find relevant case laws.
- **Case Prediction:** AI predicts the outcomes of legal cases based on historical data.
- **Document Automation:** AI simplifies document drafting and filing.
- **Client Support:** Chatbots provide legal advice and handle queries in real-time.

8. Security and Defense

AI enhances security measures and operational efficiency in defense sectors.

- **Cybersecurity:** AI systems like Darktrace detect and respond to cyber threats in real-time.
- **Surveillance:** Facial recognition and behavior analysis improve public safety.
- **Autonomous Drones:** AI-powered drones assist in reconnaissance and disaster relief.
- **Threat Detection:** AI predicts and prevents security breaches based on patterns.
- **Command Automation:** AI streamlines decision-making in complex military operations.

9. Environment and Climate

AI contributes to environmental conservation and climate change mitigation.

- **Wildlife Monitoring:** AI tools like Wildbook track endangered species using image recognition.
- **Climate Prediction:** AI models simulate climate scenarios and recommend solutions.
- **Deforestation Prevention:** Tools like Global Forest Watch use AI to monitor illegal logging activities.
- **Waste Management:** AI-powered robots sort recyclable materials efficiently.
- **Carbon Footprint Reduction:** AI helps optimize energy usage and promotes green technologies.

10. Human Resources

AI simplifies recruitment and employee management processes.

- **Resume Screening:** Tools like HireVue use AI to shortlist candidates based on job requirements.
- **Employee Engagement:** AI analyzes feedback to improve workplace culture.
- **Performance Tracking:** AI monitors productivity and recommends training programs.
- **Workforce Planning:** Predictive analytics helps in optimizing team structures.

AI is transforming industries by improving efficiency, reducing costs, and enabling innovation. As AI continues to evolve, its applications will become even more pervasive, addressing challenges and unlocking opportunities across all sectors.

3.2 APPLICATIONS OF AI IN HEALTHCARE

AI is revolutionizing healthcare by enhancing patient outcomes and simplifying administrative procedures.

3.2.1 MEDICAL IMAGING

AI systems are able to precisely identify anomalies like tumors, fractures, and infections by analyzing medical pictures, including X-rays, MRIs, and CT scans. AI is helping doctors to see inside the body better, that means doctors can figure out what's going on faster and more accurately, which is great for patients and for better diagnosis. For example, Google Health developed an AI model to detect breast cancer in mammograms. Other applications include detecting lung cancer from CT scans, identifying brain anomalies like strokes, and diagnosing retinal diseases using AI in ophthalmology.

3.2.2 DISEASE PREDICTION

AI can forecast disease outbreaks, readmissions, and the course of chronic illnesses by analyzing patient data and medical records. AI is widely used to enhance patient outcomes, reduce costs, and improve operational efficiency.

For example,

- **The HINT Model (Hospital Intelligent Notification Tool) :** Hospitals use tools like HINT to monitor patients continuously and predict the likelihood of sepsis within hours, improving response times.
- **Google's AI Model:** Google has developed AI models capable of predicting acute kidney injury (AKI), a related condition, 48 hours in advance with high accuracy.

3.2.3 PERSONALIZED MEDICINE

By using machine learning algorithms to customize treatments for each patient according to their genetic composition and medical history, medicines become more successful.

AI examines your medical history, including your genes, past medical events, and treatment outcomes. It then creates a unique plan for you. This indicates that your treatment is more effective and causes less issues for you. It helps prevent problems brought on by incorrectly given medication, much like having a personal health coach.

3.2.4 RAPID AND AFFORDABLE DRUG DEVELOPMENT

In the lab, AI behaves like a super scientist. It predicts how various compounds can cure illnesses using specific algorithms. This enables us to produce novel medicines much more quickly and affordably that allow people to receive the necessary therapies more quickly and at a lower cost of production. Following are the benefits of incorporating AI in the field of drug development:

- **Speed:** AI reduced the drug discovery timeline from years to months.
- **Cost-Effectiveness:** By prioritizing promising compounds, AI minimizes the need for costly trial-and-error experiments.
- **Precision:** AI ensures targeted drug development, increasing the likelihood of success.

Example: One of the most notable examples of AI's impact on medication research is its help in speeding up the discovery of therapies and vaccinations during the COVID-19 epidemic. Traditional medication development techniques required a large amount of time and money, which AI-powered tools greatly decreased. Benevolent AI, an AI-based drug discovery company, identified an existing drug, Baricitinib, as a potential treatment for COVID-19.

3.2.5 VIRTUAL HEALTH ASSISTANTS

Chatbots and virtual assistants driven by AI offer patients round-the-clock assistance by answering inquiries, setting up appointments, and giving medical advice.

These virtual health assistants provide various benefits like:

- **24/7 Accessibility:** Users can access medical advice anytime, reducing the burden on healthcare facilities for non-urgent queries.
- **Cost-Effective Care:** Virtual assistants lower the cost of healthcare delivery by automating initial consultations.
- **Improved Patient Engagement:** They provide personalized health insights and reminders for medications or lifestyle changes.

Example

- Babylon Health, a UK-based health technology company, offers a virtual health assistant that combines AI-driven chatbots and video consultations with doctors.
- Similar tools like **Ada Health**, **Buoy Health**, and **Microsoft Healthcare Bot** are also being widely adopted in various healthcare systems globally.

3.2.6 MANAGING HOSPITAL OPERATIONS AND RESOURCES

AI serves as a hospital manager. It assists with scheduling patient arrivals, allocating resources such as physicians and supplies, and ensuring that everything functions well. It may even predict the number of patients that may arrive in advance, allowing hospitals to make the most efficient use of their personnel and resources.

Example

- Johns Hopkins Hospital implemented an AI-powered command center to optimize its operations. The system provides a centralized view of hospital activities, using predictive analytics to anticipate bottlenecks. It predicts patient discharge times and allocates beds more efficiently, reducing waiting times for incoming patients.

3.2.7 DEVELOPMENT OF AI ENABLED HEALTH CARE DEVICES:

There are numerous AI-enabled healthcare devices that are transforming the medical industry **for monitoring, wearing, robotic surgery, imaging, personal virtual assistance, etc.:**

- Diagnostic and Monitoring Devices like IDx-DR that detects diabetic retinopathy by analyzing retinal images. Viz.ai detects strokes in CT scans and alerts healthcare teams for timely intervention.
- Wearable Devices like Apple Watch (with AI-based ECG) for tracking heart rhythm detecting atrial fibrillation (AFib), Fitbit Sense for stress management, heart rate tracking, and early detection of potential health issues, etc.

3.3 APPLICATIONS OF AI IN EDUCATION

Artificial intelligence (AI) is transforming education by allowing for personalized learning, automating administrative tasks, and increasing the accessibility and diversity of educational information. Here, we look at real-world uses of artificial intelligence in education, proving its transformational potential.

3.3.1 PERSONALIZED LEARNING

Each student possesses unique capabilities, limitations, and learning styles. Artificial intelligence programs may learn from students' study habits and provide individualized learning paths, assisting teachers to focus on all pupils at once. The main aim of Artificial Intelligence in education is not to completely replace teachers. Instead, it aims to act as helping hands for teachers as well as students. Teachers can create a study plan that is tailored to each student's needs. As AI advances, it is feasible that robots may be able to analyze students' facial expressions while learning topics, determine if they are having difficulties in learning, and make modifications to the teaching method accordingly.

Examples:

- **DreamBox Learning:** DreamBox Learning, an adaptive learning platform, employs artificial intelligence to monitor student performance in real time. It customizes mathematics lessons based on individual skills, limitations, and learning styles.

For example, if a student suffers with fractions, DreamBox adapts the curriculum to emphasize this issue.

- **Squirrel AI:** Squirrel AI, a Chinese company, uses AI to provide tailored teaching in areas such as mathematics and English. The platform's algorithms examine hundreds of data points, including time spent on each question and mistake patterns, to develop a personalized learning path for each learner.
- **Virtual Tutors:** AI-powered virtual tutors offer students additional support and tutoring, helping them understand complex concepts.

3.3.2 AUTOMATING GRADING AND FEEDBACK TASKS

In addition to tailoring the teaching process, AI software systems may verify tasks, grade tests, organize research papers, keep reports, and prepare presentations and notes, among other administrative tasks. Thus, utilizing AI to automate common operations has increased productivity in the learning environment. AI systems grade assignments and exams, providing timely feedback and freeing teachers' time for more personalized instruction. AI streamlines administrative operations, allowing educators to devote more time to teaching and less to paperwork.

Examples:

- **Gradescope:** Turnitin created Gradescope, which employs artificial intelligence to evaluate assessments. Identifying trends in student replies allows for consistent and objective grading, even for open-ended questions. For example, in large university courses, instructors can use Gradescope to evaluate hundreds of essays efficiently.
- **Google Classroom:** Google Classroom uses artificial intelligence to automate regular operations like sorting assignments, issuing reminders, and managing timetables. Teachers may utilize these technologies to save time on administrative tasks and focus more on student engagement.

3.3.3 ENHANCING ACCESSIBILITY AND INCLUSIVITY

AI tools are making education more accessible to students with diverse needs and abilities. Students with learning disabilities, such as hard of hearing, visually impaired and so on, face several problems in life. Such learners may experience a variety of obstacles when learning and studying. Furthermore, they require extra care and attention. With the use of cutting-edge AI technology, there will be new methods to communicate with these students. AI-enabled tools can be effectively trained to assist a group of students with special needs.

Examples:

- **Microsoft Immersive Reader:** Microsoft Immersive Reader assists students with reading difficulties, such as dyslexia. It uses text-to-speech AI, adjustable text sizes, and translation tools to make reading content more manageable. For instance, a student who struggles with reading comprehension can listen to text being read aloud while following along visually.

- **Seeing AI:** Seeing AI, an app developed by Microsoft, helps visually impaired students by narrating the text and surroundings in real-time. This app enables visually impaired learners to access textbooks, classroom notes, and other educational materials.

3.3.4 INTELLIGENT TUTORING SYSTEMS

Artificial intelligence chatbots/tutors are useful for learners who need extra instruction after school. The idea is to allow students to speak with the bots as they would with real professors, resulting in a personalized, one-on-one learning experience. AI-powered tutoring systems provide additional support to students outside the classroom. Currently, there are various AI-driven tutoring programs that can help students in learning the basics of mathematics, writing, and other subjects.

Examples:

- Duolingo, a popular language learning program, employs artificial intelligence to tailor lessons to the user's progress. For instance, if a learner repeatedly face difficulties with verb conjugations, the app will increase the frequency of related activities to help reinforce the topic.
- Carnegie Learning's MATHia, an intelligent tutoring system, specializes in mathematics learning. It employs AI to deliver step-by-step instructions, simulating the experience of working with a personal teacher. For example, if a learner is struggling with algebraic equations, MATHia provides pointers and alternate explanations to help reinforce comprehension.

3.3.5 LANGUAGE LEARNING AND TRANSLATION

AI-driven language tools facilitate multilingual education and improve communication.

Examples:

- Google Translate's AI-powered translation capabilities enable teachers and pupils to interact across linguistic boundaries. For example, in classes with non-native speakers, teachers can utilize Google Translate to give instructions in the students' native languages.
- ELSA Speak is an artificial intelligence tool that helps learners improve their English pronunciation. It gives real-time feedback and practice sessions that are personalized to individual needs, allowing non-native speakers to acquire confidence in their speaking skills.

3.3.6 PREDICTIVE ANALYTICS FOR ACADEMIC SUCCESS

AI systems analyze data to predict student outcomes, enabling early interventions.

Examples:

- Civitas Learning uses AI to identify students at risk of dropping out by analyzing attendance records, grades, and engagement metrics. For instance, universities use Civitas to offer tailored support services, such as tutoring or counseling, to the needy students.

- BrightBytes leverages AI to assess the impact of technology on learning outcomes. Schools can use this platform to optimize their digital investments, ensuring resources like tablets and online tools are effectively enhancing student performance.

3.3.7 MAKING THE LEARNING FUN WITH GAMING

AI can be employed in games driven educational systems, making learning more interesting and enjoyable.

Examples:

- Kahoot! utilizes artificial intelligence to produce adaptive quizzes for students of various skill levels. Teachers may utilize the platform to create interactive games that enhance classroom topics and make learning more engaging.
- Minecraft: Education Edition use AI to create interactive quizzes and simulations that respond to student input, keeping learners motivated and involved.
- Prodigy gamifies mathematics teaching by incorporating curriculum-based information into an adventure game. For example, pupils answer arithmetic problems to advance through stages, with AI tailoring the challenge to their abilities.

3.3.8 SMART CONTENT CREATION

Teachers and researchers may also use artificial intelligence and machine learning to generate new information, making teaching and learning simpler.

AI helps educators create digital classes and study materials. Tools such as Magic School AI and Eduaide. AI makes lesson preparation easier, creates evaluations, writes individualized education plans (IEPs), and does much more to modernize learning and streamline instruction.

Examples:

- **Quizlet:** Quizlet employs AI to generate flashcards and quizzes based on textbook content or user-provided material. Its Learn Mode adapts questions to help students focus on areas needing improvement.
- **Content Creation in STEM:** AI platforms like Wolfram Alpha help generate problem sets and visual explanations for complex mathematical and scientific concepts.

3.3.9 CURRICULUM DEVELOPMENT

AI facilitates the development of instructional material and curriculum design. AI assists educators in developing courses by analyzing educational data to discover patterns and gaps. This ensures that the curriculum is relevant, thorough, and connected with learning objectives by recommending revisions based on current educational requirements and standards.

Example:

- Content Technologies Inc.'s (CTI) AI algorithms generate tailored textbooks by evaluating course syllabi and student information. For example, a university

economics course may employ AI-generated textbooks that are personalized to the precise themes addressed in the class.

- ScribeSense uses artificial intelligence to scan and evaluate handwritten assignments, allowing instructors to more easily assess student work and find areas for growth.

3.3.10 INCLUSIVE CLASSROOMS USING ADAPTIVE TECHNOLOGY

AI-powered adaptive technologies promote equal learning experiences. AI-driven platforms assess students' skill levels in real time and tailor instructional content to meet individual needs. These systems adapt lessons dynamically based on student responses, providing customized pathways to help students to master concepts at their own pace.

Examples:

- Quizlet's AI-powered adaptive study regimens accommodate to individuals' various learning demands, allowing them to proceed at their own speed. For example, a student studying for a biology test might concentrate on weaker areas identified by Quizlet's algorithms.
- Otter.ai offers real-time transcription of lectures, making them more accessible to students with hearing problems. It also improves note-taking by allowing all students to access full records of classroom conversations.

3.3.11 AI FOR AUTOMATING THE ADMINISTRATIVE TASKS IN EDUCATIONAL INSTITUTIONS

AI is revolutionizing administrative tasks in educational institutions by streamlining processes, improving efficiency, and reducing manual effort. Let's have a look on the prominent examples:

- **Automated Grading :** AI-powered tools like Gradescope by Turnitin, assist teachers in grading assignments, quizzes, and exams faster and more consistently, freeing up time for instructional activities.
- **Student Enrollment and Admissions:** Many universities use AI chatbots like AdmitHub (now Mainstay) to guide prospective students through the application process, answering common questions and providing reminders about deadlines.
- **Scheduling and Timetable Management:** Edsby employs AI to optimize scheduling for classes, exams, and faculty availability, reducing conflicts and enhancing resource utilization.
- **Predictive Analytics for Student Retention:** Civitas Learning analyze student data to predict those at risk of dropping out, enabling early interventions and tailored support to improve retention rates.
- **Attendance Management:** Some institutions use AI-driven facial recognition to automate attendance tracking, ensuring accuracy while saving teachers' time.
- **Document Management:** AI systems like IBM Watson help in digitizing and organizing administrative records, making it easier to retrieve and analyze institutional data.

3.3.12 AI-POWERED ASSESSMENT TOOLS.

AI-powered assessment systems expedite evaluations and provide more detailed information on student performance.

Example:

- The PTE Academic English test from Pearson assesses speaking, writing, reading, and listening skills using artificial intelligence. Its grading method maintains objectivity and uniformity.
- Turnitin, uses artificial intelligence to assess papers for originality and provide feedback on writing quality, allowing students to improve their contributions before to final grade.

3.3.13 ENHANCED STEM EDUCATION

AI enhances STEM education by allowing for creative approaches to teaching science, technology, engineering, and mathematics.

Example:

- AI simulations Tools such as Labster offer AI-powered virtual labs, allowing students to conduct experiments without the need for physical lab equipment.
- Google's AI products, such as Blockly, offer visual coding platforms that introduce programming principles to young learners in a simple way.

3.3.14 CHALLENGES AND FUTURE DIRECTIONS FOR AI IN EDUCATION

While the potential for AI in education is enormous, issues such as data protection, ethical use, and equity must be addressed. As technology advances, the use of AI in education is anticipated to increase, resulting in a more inclusive and effective learning ecosystem.

Education is no longer limited by traditional constraints thanks to the use of AI which provides learners with previously inconceivable tools and chances.

3.4 APPLICATION OF AI IN AGRICULTURE

Artificial Intelligence (AI) is becoming increasingly prevalent in agriculture. T maximize productivity, the industry must address difficulties such as poor soil treatmen disease and pest infestation, massive data requirements, low output, and a knowledge ga between farmers and technology. AI in agriculture focuses on flexibility, high performanc accuracy, and cost-effectiveness.

At this crucial moment, Artificial Intelligence presents unparalleled prospects f agriculture. AI has a wide-ranging impact, from improving food output and quality optimizing resource consumption. Whether it's assessing land use with high-precisio satellite imagery or detecting crop diseases through real-time monitoring, AI applicatio are gaining attraction worldwide. This wave of technology is not only gaining attracti in agri-tech, but it is also attracting investments to drive innovation and growth.

Let's chart out some of the important applications of AI in agriculture:

3.4.1 DISEASE AND PEST MANAGEMENT

Disease control is crucial for achieving optimal agricultural yields. Diseases of plants and animals greatly hinder crop growth. Diseases that affect plants and animals can be caused by a variety of variables, including genetics, soil type, weather patterns, wind, and temperature. Managing the consequences of certain illnesses can be difficult, especially in large-scale farming, due to their unpredictable nature and other reasons. To effectively control illnesses and minimize losses, farmers should implement AI.

AI aids in early detection of diseases and pest infestations, enabling timely intervention and reducing crop losses.

Examples:

- **Plantix:** This app can identify plants from images, detect diseases, and suggest treatments. It also has a community where plant lovers can share information.
- **AgroAI:** This app can calculate the severity of diseases and evaluate soil fertility.
- Other apps that can help detect plant diseases include: Agrivi, FarmRite, and CropIn.
- **Xarvio's Digital Farming Solutions:** AI models predict pest outbreaks and recommend preventive measures.

3.4.2 AUTOMATED WEED CONTROL SYSTEMS

Computer vision can be used to identify invasive plant species and weeds, just as it can identify diseases and pests. Computer vision uses the size, shape, and color of leaves in conjunction with machine learning to differentiate crops from weeds. Robots that perform robotic process automation (RPA) tasks, like automatic weeding, can be programmed using such solutions.

Example:

- The 'See & Spray' technology, was created by Blue River Technology, which is now a part of John Deere. It uses AI algorithms and high-resolution cameras to detect weeds among crops. Compared to conventional methods, this technology reduces herbicide usage by up to 90% by enabling precise application.

3.4.3 PRECISION AGRICULTURE

Precision agriculture uses AI-driven tools to analyze various factors such as soil conditions, weather patterns, and crop health, enabling farmers to make data-driven decisions.

Examples:

- **Blue River Technology:** Offers AI-powered tools that enable precision weeding, saving costs and reducing environmental impact.
- **eAgronom in Europe:** Analyzes satellite imagery and soil data to recommend optimal planting times and nutrient requirements.

3.4.4 CROP MANAGEMENT AND MONITORING

Drones with AI capabilities and satellite imagery are revolutionizing crop monitoring by offering real-time information on pest infestations, crop health, and moisture levels.

Examples:

- AI algorithms and high-resolution aerial imagery are used by agritech startups like Taranis to detect diseases and pests early on.
- In Israel, CropX uses AI and IoT sensors to create comprehensive maps of soil moisture, which aid farmers in improving irrigation techniques.
- **Hortau (Canada) :** Uses AI to track plant stress so that yields can be increased in a timely manner.

3.4.5 PREDICTIVE ANALYSIS FOR WEATHER AND YIELD

Artificial intelligence (AI) systems use both historical and current data to forecast weather, pest outbreaks, and crop yields with high accuracy.

Examples:

- Using predictive analytics, IBM's Watson Decision Platform for Agriculture provides farmers with information on when to plant, when to harvest, and how to control pests.
- The Climate Corporation assists farmers in making efficient plans by using AI to generate hyper-local weather forecasts and field-specific recommendations.
- In India, Microsoft AI reduces losses by up to 20% by collaborating with local governments to forecast pest infestations in cotton crops.

3.4.6 INTELLIGENT IRRIGATION SYSTEMS

For agriculture to be sustainable, water management is essential. Smart irrigation systems powered by AI maximize water use while cutting waste and enhancing crop health.

Examples:

- Netafim's Precision Irrigation System analyzes soil conditions and weather data using AI algorithms to automate irrigation control.
- Water resources are conserved by AI-powered systems from HydroPoint Data Systems (USA) that track water consumption and stop excessive irrigation.
- **FarmBot:** An open-source artificial intelligence platform that uses weather forecasts and plant requirements to automate watering.

3.4.7 DRONE-ASSISTED AERIAL SURVEILLANCE

Drones, also known as Unmanned Aerial Vehicles (UAVs), are used to capture aerial imagery, videos, and data over large agricultural areas. Equipped with sensors like RGB cameras, multispectral sensors, and LiDAR, drones provide detailed insights into various aspects of farming. By combining drones with AI, machine learning, and advanced imaging systems, farmers can gather real-time, high-resolution data about their fields, enabling

informed decision-making. This approach improves productivity, optimizes resource use, and promotes sustainable farming practices.

Examples:

- **Taranis (Israel)** : Uses drones to capture ultra-high-resolution images for crop health monitoring, identifying diseases, and recommending solutions.
- **FarmShots (USA)** : Employs drones to create NDVI maps, helping farmers detect stress caused by pests, diseases, or drought.
- **XAG (China)** : Develops drones for precision spraying and monitoring, enhancing efficiency in rice and wheat farming.

3.4.8 LIVESTOCK MANAGEMENT

Livestock management include disease prediction, feeding schedule optimization, and animal health monitoring.

Examples:

- **Cainthus (Ireland)** : Monitors the eating, drinking, and health habits of dairy cattle using computer vision.
- The Ida System from Connecterra examines cow behavior to enhance milk production and identify health problems early.
- Allflex Livestock Intelligence creates wearable AI tools for monitoring health and movement indicators in animals.

3.4.9 SUPPLY CHAIN OPTIMIZATION AND DEMAND FORECASTING

Supply chain optimization in agriculture involves using technology, including AI, to streamline and improve the processes of producing, storing, transporting, and delivering agricultural products. This ensures that goods move efficiently from farms to consumers while minimizing waste, reducing costs, and enhancing profitability.

Examples:

- **Zymergen**: Uses AI to optimize agricultural inputs like fertilizers and seeds, reducing costs.
- **Descartes Labs (USA)** : Analyzes satellite data to forecast crop yields and market prices, helping farmers make better decisions.
- **Farm to Market Alliance (FTMA)** : Employs AI to connect farmers with buyers, ensuring fair prices and minimizing waste.

3.4.10 AUTONOMOUS FARMING EQUIPMENT

AI is at the core of autonomous tractors, harvesters, and drones, reducing labor dependency and increasing operational efficiency.

Examples:

- **Case IH and New Holland (USA)** : Develop autonomous tractors that use AI for precision plowing and planting.

- **Harvest CROO Robotics:** Builds robotic strawberry pickers capable of harvesting 8 acres per day, addressing labor shortages.
- **DJI Agriculture Drones:** Equipped with AI for spraying, seeding, and monitoring vast areas efficiently.

3.4.11 WEATHER AND PRICE FORECASTING

Farmers are finding it challenging to make the best decisions regarding harvesting, seeding, and soil preparation because of climate change. However, farmers can plan for the type of crop to grow, the seeds to sow, and the crop harvesting process by using AI. weather forecasting to get information on weather analysis. Price forecasting helps farmers make the most profit by giving them a better idea of the price of crops for the upcoming weeks.

Examples:

- AgNext (India) uses AI to analyze real-time market trends, weather data, and crop quality to provide price forecasts for farmers and traders.
- Fasal (India) an AI-based agritech platform, integrates IoT and AI to provide farmers with price predictions for fruits and vegetables. It combines weather data and market trends to offer precise recommendations.
- Microsoft AI partnered with the Indian government to develop an AI platform that predicts weather conditions and pest outbreaks for specific crops like cotton.
- IBM's Watson Decision Platform for Agriculture analyzes real-time weather data, historical climate patterns, and soil conditions to offer actionable insights for farmers.

3.4.12 FARMER TRAINING AND EDUCATION

AI-powered platforms provide farmers with real-time learning resources and advisory services.

Examples:

- **Digital Green (India) :** Uses AI to tailor video-based training content for rural farmers, improving adoption rates of best practices.
- **FarmBeats by Microsoft:** Leverages AI and IoT to offer actionable insights to farmers through user-friendly platforms.
- **Pinduoduo (China) :** Incorporates AI in e-commerce platforms to train farmers in efficient marketing and distribution.

3.4.13 CHALLENGES AND FUTURE OF AI IN AGRICULTURE

Many people believe AI is only relevant to digital tasks and has no relevance to actual farming work. Usually, this assumption stems from a lack of knowledge about AI tools. The adoption of AI in the agricultural sector is sluggish because most people, particularly those in non-technical fields, don't fully understand how AI in agricultural biotechnology operates. Many farmers are more accustomed to using traditional methods, despite the

fact that agriculture has seen numerous advancements over its lengthy history. It's unlikely that majority of farmers have experienced working on AI-related projects.

While AI offers numerous benefits, challenges like high costs, lack of digital literacy, and data privacy concerns must be addressed. Efforts to make AI accessible to small holder farmers and ensure ethical data use are crucial for widespread adoption.

AI in agriculture is no longer a futuristic concept but a transformative force addressing some of the industry's most pressing challenges. From precision farming to climate resilience, AI is enhancing efficiency, sustainability, and profitability. As technology continues to evolve, collaborative efforts between governments, private sectors, and farmers will ensure AI's full potential is realized in agriculture.

3.5 PROBLEM SOLVING

Problem-solving in **Artificial Intelligence (AI)** refers to the process of designing systems or algorithms that can find solutions to complex problems by mimicking human reasoning or decision-making. It involves identifying challenges, analyzing situations, and applying strategies to find effective solutions.

Steps in AI Problem-Solving

1. **Define the Problem:** Clearly specify the initial state, the goal state, and the possible actions.

   Example: A maze-solving robot starts at the maze's entrance (initial state) and needs to reach the exit (goal state).

2. **Represent the Problem:** Use a structured format, like a state-space representation, where every possible situation is a "state."

   Example: In a maze, each position the robot can occupy is a state.

3. **Search for Solutions:** AI uses search algorithms to explore the possible states and find a path to the goal. Search can be uninformed (random exploration) or informed (guided by heuristics or rules).

4. **Evaluate the Solution:** Once a solution is found, evaluate its efficiency (e.g., is it the shortest or fastest path?).

3.5.1 DEFINING THE PROBLEM:

Suppose we start with the problem statement "Solve a maze." For a machine, this is an incomplete description of the task we want to accomplish. To build a program that could "solve a maze," we would first need to define the starting position in the maze, the rules that specify valid moves, and the condition that represents reaching the exit.

For the problem "Solve a maze," it is relatively straightforward to provide a formal and complete problem description. The starting position can be represented as a specific cell in the maze (e.g., the top-left corner). The goal can be described as reaching the exit cell (e.g., the bottom-right corner). The legal moves, such as moving up, down, left, or right, provide the way to transition from the starting position to the exit while avoiding walls or invalid paths.

However, if we were to write rules for every possible maze configuration, the number of rules would be immense since there would need to be separate rules for each possible layout of the maze. This would present two significant challenges:

1. It would be impractical for a human to create such an exhaustive set of rules— it would take too much time and be prone to errors.
2. A program would struggle to efficiently process and manage such an enormous rule set.

To overcome these issues, the rules for valid moves should be described in a more generalized and abstract manner. This can be achieved by representing the problem as a **State Space Search.**

3.5.2. REPRESENT THE PROBLEM AS A STATE SPACE SEARCH

- **Initial State:** The starting point of the problem.
  Example: In a chess game, the initial state is the standard arrangement of pieces on the board.
- **Goal State:** The desired outcome or solution to the problem.
  Example: In a maze, the goal state is reaching the exit.
- **State Space:** A set of all possible configurations (or "states") that the problem can take during its resolution. Each state represents a snapshot of the problem at a specific point. Example: In a puzzle game, the state space is all possible arrangements of the puzzle pieces.
- **Actions (Operators) :** The set of valid actions or operations that can transition the system from one state to another.
  Example: In a robot navigation problem, actions might include "move forward," "turn left," or "turn right".
- **Path:** A sequence of states connected by actions that lead from the initial state to the goal state.
  Example: In a navigation problem, a path could be a series of directions to reach a destination.
- **Constraints:** Rules or restrictions on the states or actions.
  Example: In Sudoku, numbers must not repeat in rows, columns, or blocks.

Representation of "Solve the Maze" problem as a State Space Search

Problem: Solve a Maze

- **Initial State:** Starting point of the maze (e.g., top-left corner).
- **Goal State:** Exit of the maze (e.g., bottom-right corner).
- **State Space:** All possible positions in the maze.
- **Actions:** Move up, down, left, or right (within the boundaries of the maze).
- **Constraints:**

- The robot cannot move through walls.
- The robot must remain within the maze's boundaries.

Representation of "Tic-Tac-Toe" problem as a State Space Search

- Tic-tac-toe (also known as Xs and Os) is a game for two players, X and O, who take turns marking the spaces in a 3x3 grid. The player who succeeds in placing three of their marks in a horizontal, vertical, or diagonal row wins the game
- **Initial State:** Board position of 3x3 matrix with no Xs and 0s.
- **Goal State:** Board position of 3x3 matrix with either three Xs or 0s in a horizontal, vertical, or diagonal row.
- **State Space:** All possible board position of 3x3 matrix with Xs and 0s.. Since each cell can be X, O, or empty, there are 3^9=19,683 possible configurations, but only a subset is valid based on game rules.
- **Actions:** Putting 0s or Xs in vacant positions alternatively. Actions depend on the current state of the board.
- **Constraints:**
  (a) Players alternate turns.
  (b) A cell can only be marked if it is empty.
  (c) The game ends when either A player wins by aligning three marks, or the board is full (draw).

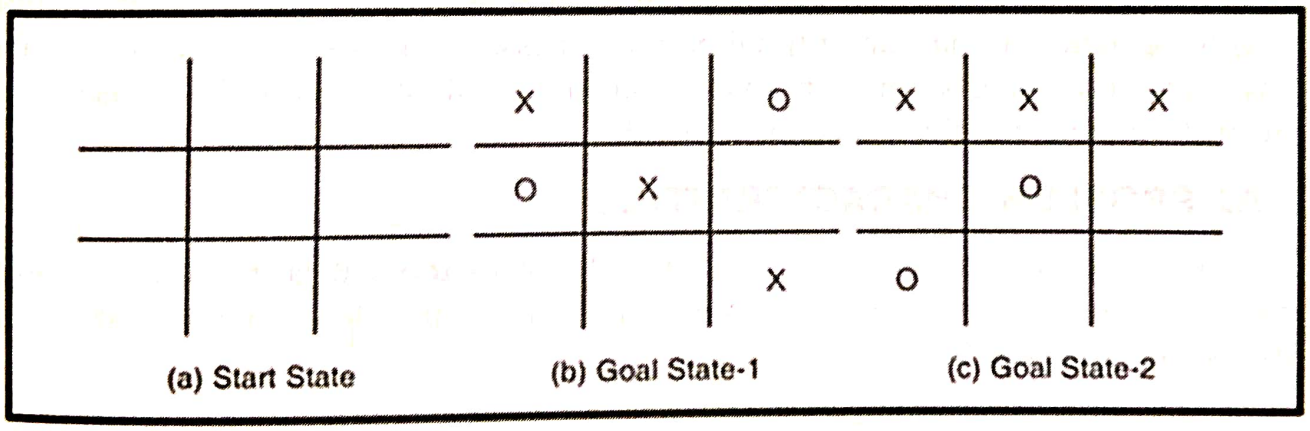


Fig. 3.2: State Space representation of 'Tic-Tac-Toe'.

3.5.3. SEARCH FOR SOLUTIONS

The problem can then be solved by applying the rules when combined with a suitable strategy to travel through the problem space until a path from a starting state to a goal is identified. This technique is referred to as **"Search". Search** is a fundamental technique in AI used to find solutions to problems by exploring all possible states systematically. When solving a problem, AI represents it as a **state space** that includes all potential configurations of the problem. The task is to navigate this space to find a path from the **initial state** (starting point) to the **goal state** (desired solution).

Search techniques allow AI systems to navigate large and complex state spaces. They Solve problems like puzzles, planning, navigation, and optimization. These techniques find efficient solutions even when the problem has multiple possible paths. Search is the backbone of problem-solving in AI, enabling machines to reason and make decisions effectively. The search techniques can be classified in two major categories:

- **Uninformed Search:** It does not use any prior knowledge about the problem.

Examples:

- **Breadth-First Search (BFS) :** Explores all states at one depth level before moving deeper.
- **Depth-First Search (DFS) :** Explores as far as possible along one branch before backtracking.
- **Informed Search:** It uses problem-specific knowledge (heuristics) to guide the search more efficiently.

Examples:

- **Greedy Best-First Search:** Prioritizes states closer to the goal based on a heuristic.
- A* Search: It combines the cost of reaching a state and the estimated cost to reach the goal.

3.5.4. EVALUATE THE SOLUTION:

Evaluating a solution involves checking if it meets the goal, satisfies constraints, and performs efficiently in terms of time, space, and other resources. In AI, the evaluation is often tied to specific criteria like optimality, robustness, scalability, and practical usability. By examining these aspects, one can ensure that the solution is not only correct but also efficient and suitable for the problem at hand.

3.5.5 AI PROBLEM CHARACTERISTICS

To choose the most appropriate problem-solving method, it is important to analyze the problem along several key dimensions, known as problem characteristics. These dimensions are outlined below:

I. Is the problem decomposable?

A large and complex problem can often be solved more easily if it can be broken down into smaller, manageable sub-problems.

II. Can solution steps be ignored or undone if they turn out to be unwise?

Problems can be classified into three categories:

- **Ignorable:** Solution steps can be ignored, and these problems can be solved using a simple control structure without needing to backtrack. An example of an ignorable problem is Theorem Proving.
- **Recoverable:** Solution steps can be undone, requiring a control strategy that allows for backtracking. An example of a recoverable problem is the 8-Puzzle.

- **Irrecoverable:** Solution steps cannot be undone, so these problems require a system that carefully considers each decision and avoids errors.

III. Is the problem universe predictable?

Problems can be classified into two types:

- **Certain Outcome:** These problems have predictable results. Examples include the 8-Puzzle and Water Jug problems.
- **Uncertain Outcome:** These problems have unpredictable outcomes. Examples include Playing Cards and Chess.

IV. Is the desired solution absolute or relative?

- **Absolute Solution:** Once a solution is found, no further solutions need to be considered.
- **Relative Solution:** After finding a solution, other solutions must be explored to determine which one is the best.

V. Is the desired solution a state or a path to a state?

- **Solution as a State:** In problems like Natural Language Processing, where the focus is on finding the interpretation of a statement, the solution is a specific state, and the steps taken to reach it do not matter.
- **Solution as a Path to a State:** In problems like the Water Jug Problem, the goal is not just the final outcome (e.g., filling the jug) but also the steps taken to achieve that outcome. The solution in such cases is a path to the final state.

VI. What is the role of knowledge?

- **Knowledge-Based:** AI uses predefined, explicit knowledge (e.g., expert systems, diagnostic systems).
- **Learning-Based:** AI learns knowledge from data (e.g., image classification, recommendation systems).
- **Hybrid:** A combination of both knowledge and learning (e.g., autonomous vehicles, healthcare diagnostics).

VII. Does the solution require human interaction?

- **Human-In-the-Loop (HITL) :** Humans guide or validate AI decisions (e.g., medical diagnosis, recommendation systems).
- **Human-Computer Interaction (HCI) :** Focus on intuitive user interfaces and communication (e.g., speech recognition, chatbots).
- **Fully Automated:** AI operates autonomously without human interaction (e.g., stock trading algorithms, industrial robots).

The table below shows the characteristics of problem playing chess.

Playing Chess:

Characteristics	Answer	Description
Is problem decomposable?	No	Problem can't b broken down into sub-problems
Can be solution be ignored or undone ?	No	Previous steps can't be undone
Is problem universe predictable?	No	There are multiple possible outcomes
Is the solution absolute or relative?	Absolute	Once you get the solution, other possible solutions don't matter
Is the solution state or path to a state?	Path	The solution is a path to the goal state
What's the role of knowledge?	.....	Helps to filter the search but not absolutely necessary
Does the task needs human intervention?	Yes	Additioal assistance is required to make a move against the AI

The table below shows the characteristics of 'Water Jug Problem'

Characteristics	Answer	Description
Is problem decomposable?	No	Problem can't b broken down into sub-problems
Can be solution be ignored or undone ?	Yes	Previous steps can't be undone
Is problem universe predictable?	Yes	There are multiple possible outcomes
Is the solution absolute or relative?	Absolute	Once you get the solution, other possible solutions don't matter
Is the solution state or path to a state?	Path	The solution is a path to the goal state
What's the role of knowledge?	.....	Helps to filter the search but not absolutely necessary
Does the task needs human intervention?	Yes	Additioal assistance is required to make a move against the AI

3.6 AGENTS AND ENVIRONMENT

An AI agent is a computational system designed to perceive its environment, reason about the situation, and take actions to achieve a specific goal or maximize utility. The agent operates autonomously and uses sensors to gather information from the environment and effectors to act on that environment. In essence, an AI agent is an entity that interacts with its environment and performs actions to achieve its objectives. An AI system can be defined as the study of the rational agent and its environment. The agents sense the environment through sensors and act on their environment through actuators. These agents can range from simple programs to complex autonomous systems, like robots or self-driving cars.

3.6.1 AI AGENTS

An AI agent is something that acts based on the information it receives from its environment. Human agents have sensory organs such as ears, eyes, tongue, and nose, as well as corresponding organs like skin, hands, and legs. Robotic agents use sensors such as cameras and infrared range finders, as well as motors and actuators, to perform actions. An AI agent can have mental properties such as knowledge, belief, intention, etc. Here are the main four rules that all AI agents must follow:

Rule 1: An AI agent must be able to perceive its environment.

Rule 2: Environmental observations must be used to make decisions.

Rule 3: The decisions should lead to action.

Rule 4: The actions done by the AI agent must be reasonable. Rational acts are those that maximize performance and produce the most favorable results.

3.6.1.1 Interaction of Agents with Environment:

There are mainly two ways the agents interact with the environment, such as **perception** and **action**. The agents interact with the environment in two ways:

1. **Perception:** Perception is a passive interaction, where the agent gains information about the environment without changing the environment. The sensors of the robot help it to gain information about the surroundings without affecting the surrounding. Hence, gaining information through sensors is called perception.
2. **Action:** Action is an active interaction when the environment is changed. When the robot moves an obstacle using its arm, it is called an action as the environment is changed. The arm of the robot is called an "Effector" as it performs the action. The interaction of the Agent with the Environment is through Sensors and Effectors.

Consider the example of a chatbot which is a virtual assistant. When it reads and understands the meaning of a user's messages, it is called perception. And when it replies to the user after analyzing the user's message, it is called the action.

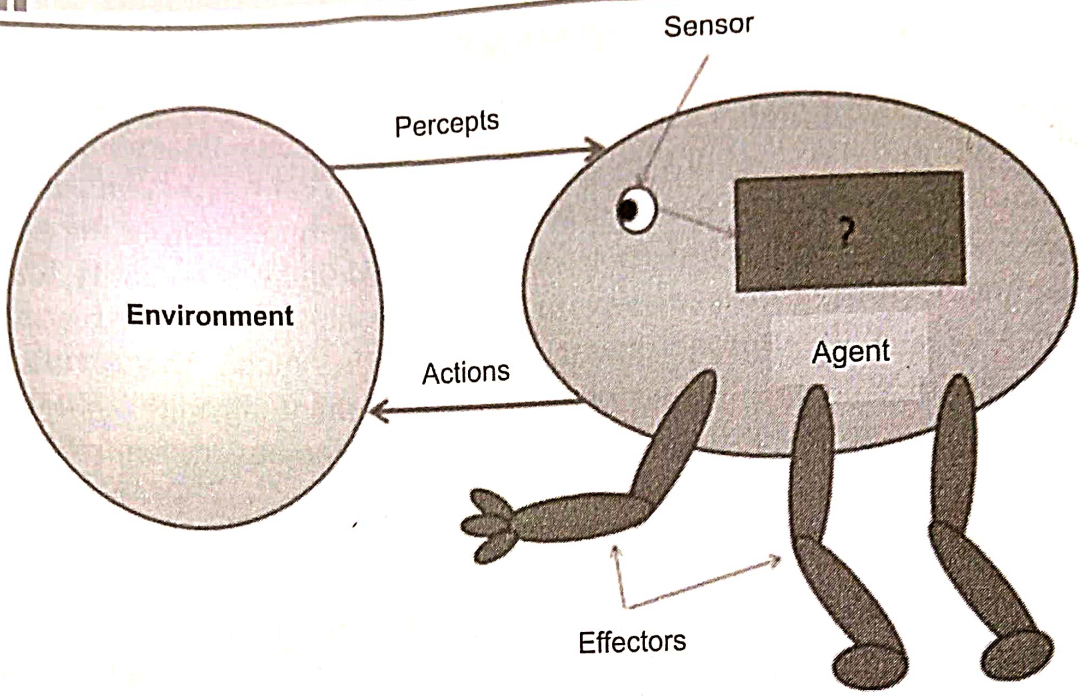


Fig. 3.3: The Agents Act in Artificial Intelligence

3.6.1.2 Structure of AI Agents

The task of AI is to design an agent program which i implements the agent function. The structure of an intelligent agent is a combination of architecture and agent program. It can be viewed as:

- **Agent** = Architecture + Agent program

  The structure of an intelligent agent is a combination of the **architecture** (the platform or hardware the agent runs on) and the **agent program** (software that implements the logic of the agent).

- **Architecture**: Architecture is machinery that an AI agent executes on. The physical platform or hardware on which the agent operates. A self-driving car uses an architecture consisting of cameras, LIDAR sensors, GPS, a computer, and actuators (steering, braking, and acceleration systems).

- **Agent Function:** Agent function F maps percept history (P*), which includes all past perceptions, to an action (A). Example: In a self-driving car, the function maps the car's sensory input (road position, nearby vehicles, and traffic signs) to an action (steering left, slowing down, or accelerating).

  $F : P \longrightarrow A$

- **Agent program:** Agent program is an implementation of agent function F. An agent program executes on the physical architecture to produce function f.It processes input from the sensors, interprets it, and determines the next action. Example: The self-driving car's program evaluates road conditions, predicts the behavior of other vehicles, and decides whether to stop, move, or turn.

Example of Self-Driving Car Agent

Component	Description
Architecture	Sensors (Cameras, LIDAR, GPS), on-board computer, actuators (steering, brakes, acceleration).
Agent Function F	Maps sensory input to driving actions. Example: If obstacle detected ahead ! Apply brakes.
Agent Program	Implements algorithms for perception (e.g., image recognition), decision-making (e.g., path planning), and control (e.g., steering logic).

How the Agent Works

1. *Percept History (P)* : The agent observes the environment using sensors like:
   - Lane markings on the road.
   - Positions of surrounding vehicles.
   - Traffic lights or signs.
2. **Agent Function (F)** : Maps sensory input to an action:
   - **Input:** A red light detected.
   - **Action:** Apply brakes.
3. **Agent Program:**
   - Evaluates the sensory data.
   - Implements driving strategies based on rules or learned behaviors (e.g., slow down in school zones or make left turns when the lane is clear).

Some other examples are presented below :

Agent Type	Architecture	Agent Function	Agent Program
Chess-Playing Agent	Computer (Processor, Memory)	Maps board positions to moves. **Example:** If checkmate possible → Move Queen to the attacking position.	Implements algorithms like Minimax or Alpha-Beta pruning.
Vacuum Cleaner Agent	Robot with suction mechanism and wheels	Maps dirty locations to movement. **Example:** If dirt detected → Move to location and clean.	Implements simple reflex-based logic or model-based reasoning.
Voice Assistant Agent	Microphone, speaker, internet connection	Maps user voice commands to actions. **Example:** If "play music" → Open music app and play song.	Natural language processing (NLP) and execution of corresponding commands.

3.6.1.3 Types of AI Agents:

AI agents are classified based on their capabilities, complexity, and how they interact with the environment. Some common types are:

1. **Simple Reflex Agents:** These are the simplest forms of agents. These agents select actions based on the current percept, using predefined rules or conditions. They do not consider the history of past actions. They have simple, reactive behavior and use condition-action rules (if-then rules). They do not have memory or planning capabilities. If we were to look at simple reflex agents and their interaction with their environment, sensors, it would look something like the image below.

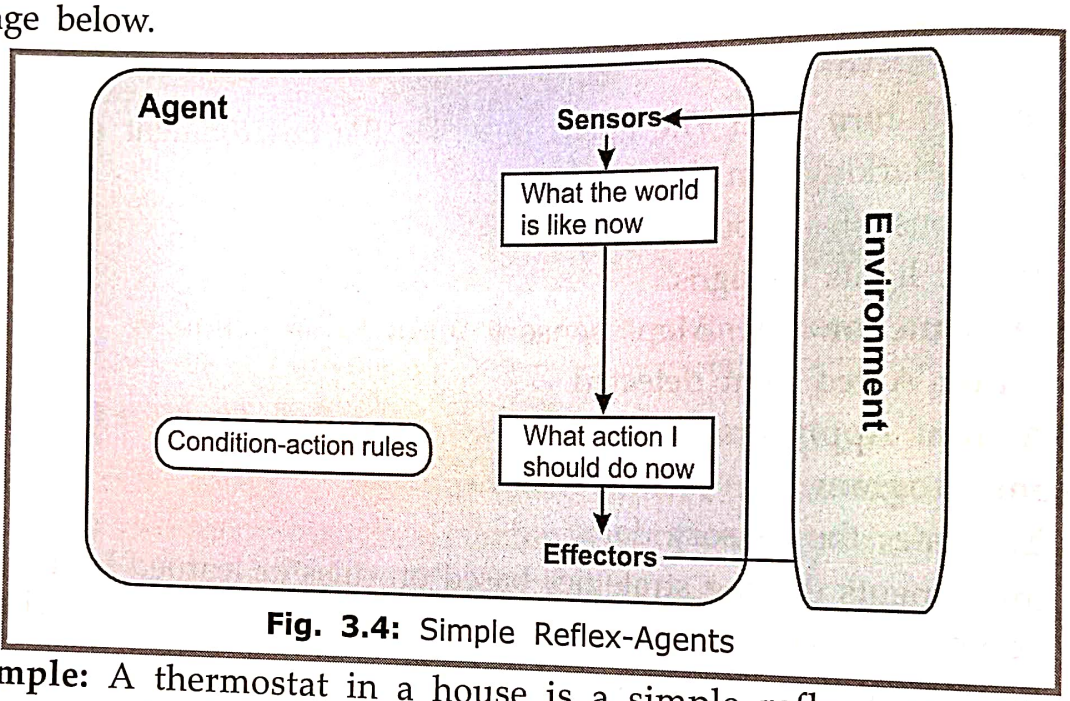


Fig. 3.4: Simple Reflex-Agents

Example: A thermostat in a house is a simple reflex agent. It adjusts the temperature based on the current readings from the thermostat (e.g., if the temperature is below 20°C, it turns the heating on).

2. **Model-Based Reflex Agents:** These agents extend simple reflex agents by maintaining an internal model of the world, which helps them handle partial

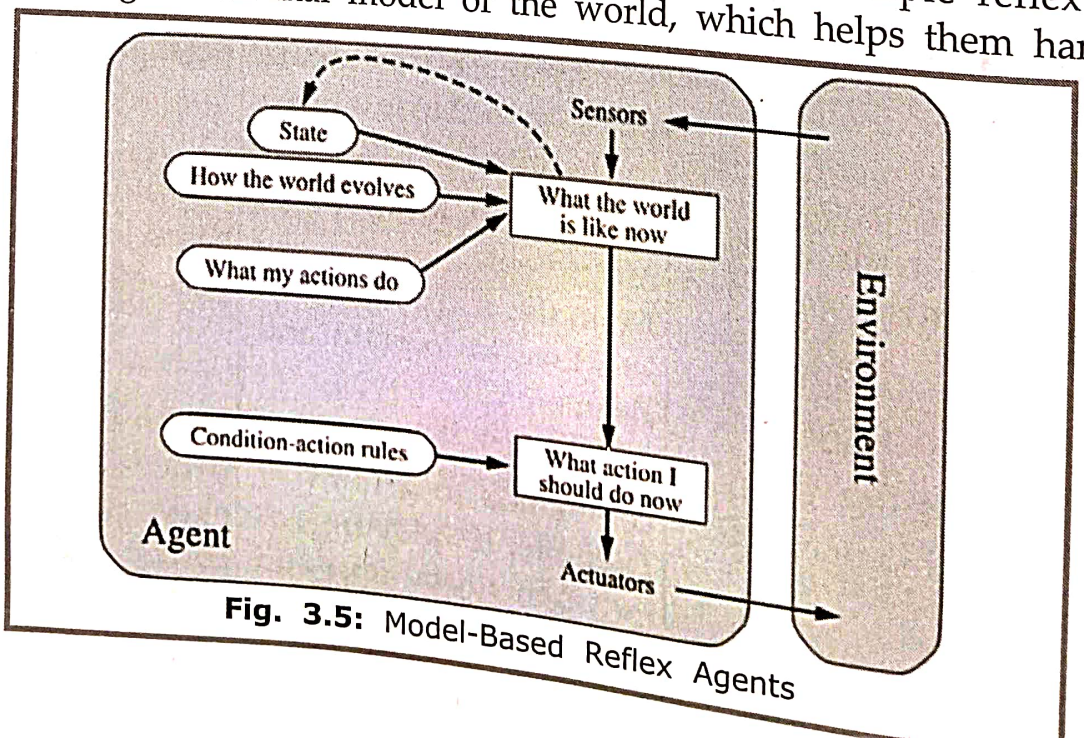


Fig. 3.5: Model-Based Reflex Agents

knowledge and deal with changing environments. They maintain an internal model or state of the environment that can handle partially observable environments and can make decisions based on past experiences. The agents have a model representing the knowledge of the world. Actions are performed based on this model. For the agent state to be updated, the following information is needed:

- How the world evolves.
- The effect of the agent's actions on the world.

Example: A robot vacuum cleaner, which builds a map of the room as it moves around, allowing it to avoid obstacles and clean efficiently.

3. **Goal-Based Agents:** These agents take actions to achieve specific goals, using a more sophisticated reasoning process. They may involve searching for solutions or planning sequences of actions. They can reason and plan ahead to achieve goals and are capable of handling complex tasks. They can also adapt to changing environments by adjusting their goals. In some cases, these types of agents have to evaluate a long sequence of actions so to know whether a goal will be achieved or not. Different scenarios have to be considered, and this process is known as searching and planning and it makes an agent be proactive.

   Example: A chess-playing AI that evaluates various moves and their outcomes to decide the best move that leads to victory.

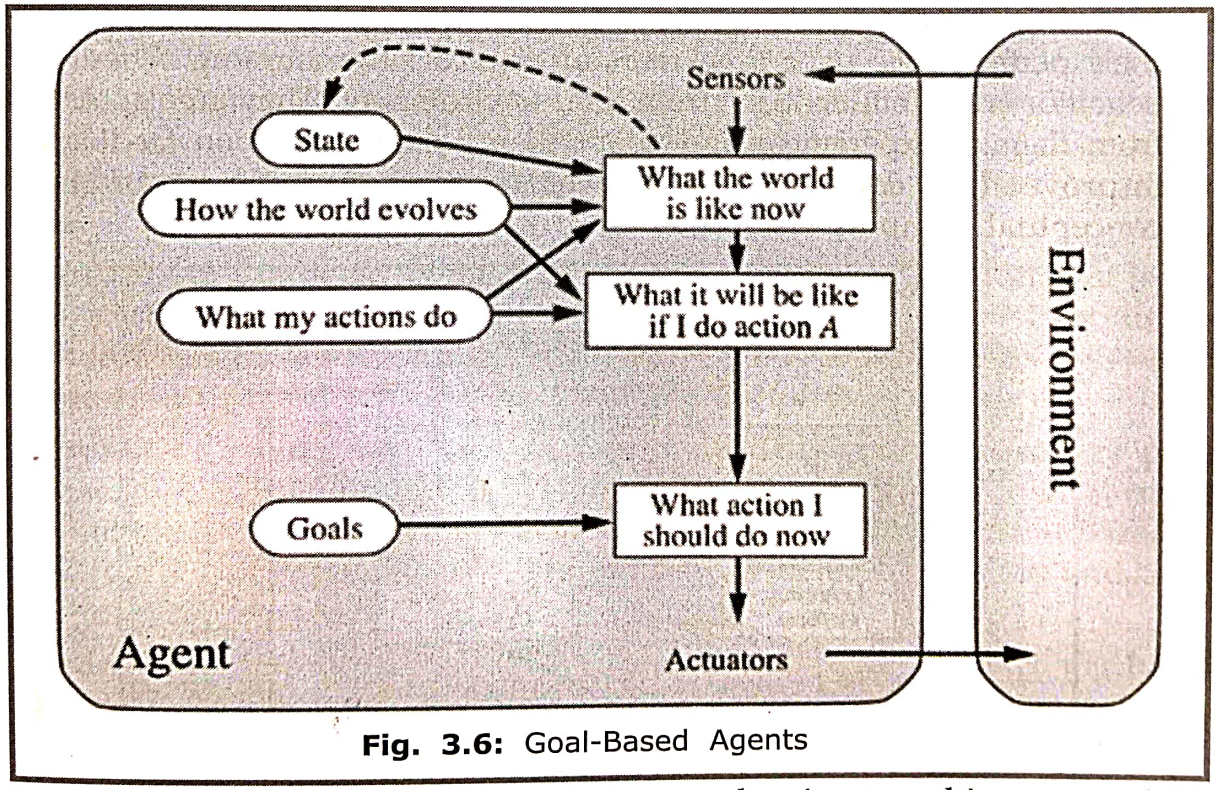


Fig. 3.6: Goal-Based Agents

4. **Utility-Based Agents:** These agents not only aim to achieve a goal but also maximize a measure of satisfaction or utility. They evaluate different possible actions based on a utility function and select the one with the highest utility. They make decisions based on maximizing a predefined utility function and can

handle multiple, conflicting goals by weighing the trade-offs. They can adapt dynamically to optimize performance.

Example: An autonomous car that not only aims to reach a destination but also maximizes safety and minimizes travel time by considering different routes.

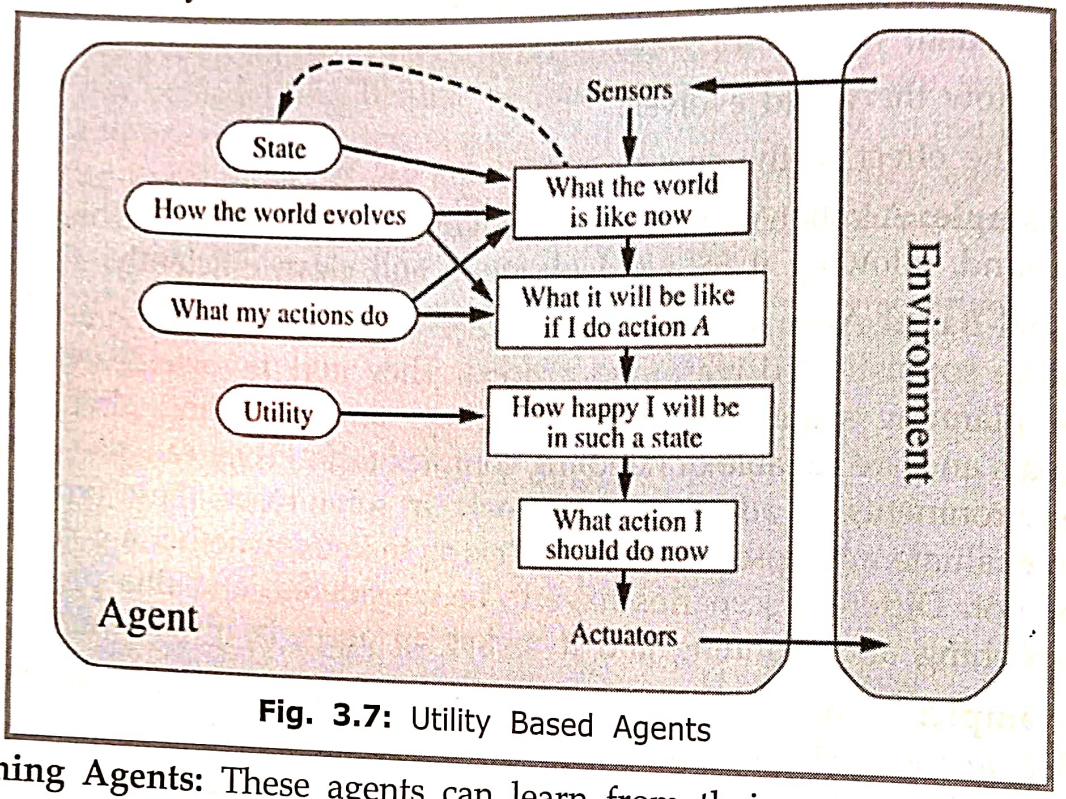


Fig. 3.7: Utility Based Agents

5. **Learning Agents:** These agents can learn from their experiences and improve their performance over time without explicit programming. They modify their behavior based on feedback from their environment. They are capable of learning from data or experiences and adapt their actions based on feedback. They can improve their decision-making over time.. It is made up of the following conceptual components:

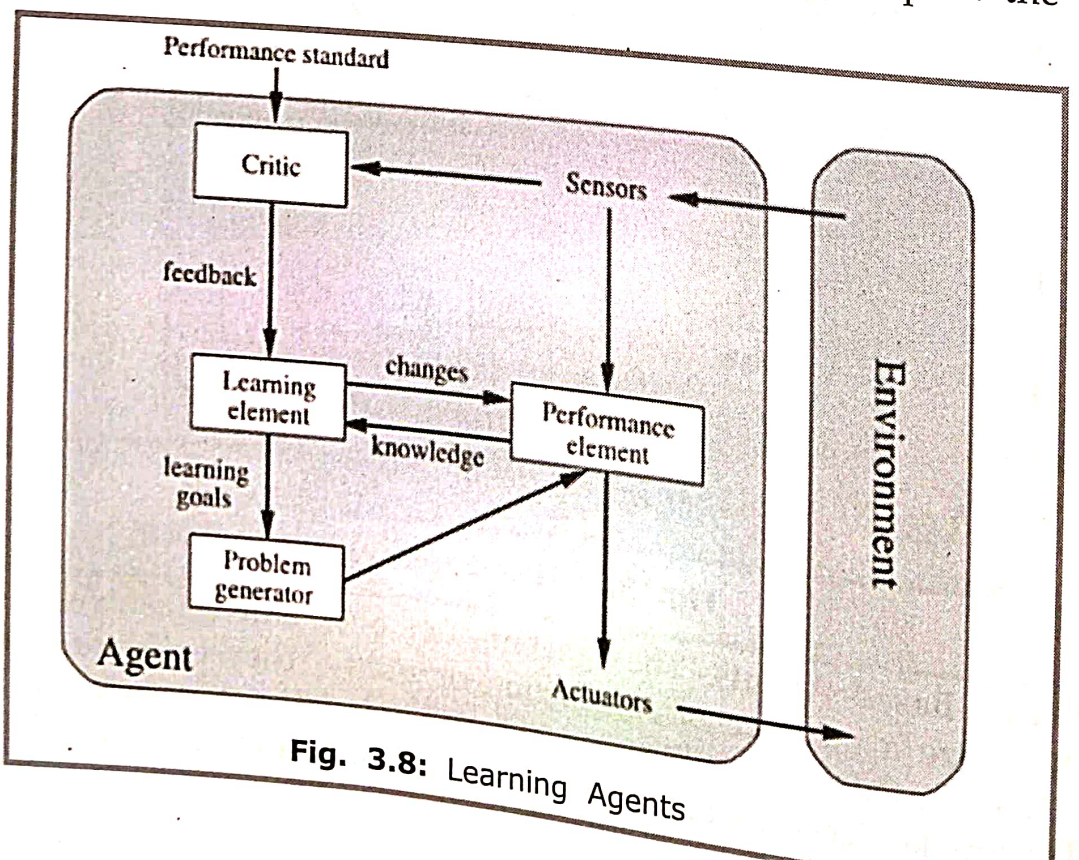


Fig. 3.8: Learning Agents

- **Learning element:** This is the element that is responsible for making improvements by learning from the environment.
- **Critic:** The learning element has to receive feedback from the critic which is the component that tells how well the element is doing with respect to a particular set performance standard.
- **Performance Element:** This is the component that selects the external action.
- **Problem Generator:** This is the component that suggests actions that may create new experiences.

Example: A recommendation system that learns user preferences over time, such as Netflix recommending shows based on past watching patterns.

3.6.1.4 Role of Agents in AI

Agents play a central role in Artificial Intelligence (AI) as they are the entities that perceive their environment, reason about it, and take actions to achieve specific goals. They form the foundation of intelligent systems by automating tasks, solving problems, and making decisions in dynamic and complex environments. Key roles of Agents in AI are:

- **Autonomous Decision-Making:** Agents act independently, making decisions without constant human intervention. They process data from the environment, analyze it, and determine the best actions to achieve their objectives. For example, a self-driving car decides when to accelerate, brake, or turn based on road conditions and traffic.
- **Problem Solvers:** AI agents are designed to solve problems by analyzing situations, exploring possible solutions, and executing the most suitable one. They can handle complex scenarios that require logical reasoning and planning. For example, a chess-playing AI evaluates potential moves and selects the best one to win the game.
- **Adaptation to Changes:** Agents adapt to changes in their environment by learning from past experiences or reacting dynamically. This makes them suitable for tasks in unpredictable settings. For example, a recommendation system adjusts its suggestions based on a user's changing preferences over time.
- **Environment Interaction:** Agents gather information using sensors or perception mechanisms, enabling them to understand and interact with their surroundings. They act through effectors or actuators to modify the environment based on their goals. For example, a robotic vacuum cleaner detects obstacles and adjusts its path to clean efficiently.
- **Learning and Improvement:** Many agents incorporate learning mechanisms to improve their performance. By analyzing feedback and outcomes, they refine their decision-making processes over time. For example, a virtual assistant learns a user's habits and provides more personalized responses over time.
- **Collaboration and Communication:** Agents often work in teams, communicating and coordinating with other agents or systems to achieve shared goals. This is common in multi-agent systems where tasks are distributed among multiple

agents. For example, delivery drones collaborating to optimize package delivery routes.

- **Optimization and Efficiency:** Agents optimize resources and processes to achieve goals in the most efficient manner. They can analyze multiple factors and constraints to find the best possible solutions. For example, AI managing supply chains optimizes inventory levels and delivery schedules.
- **Bridge between Humans and Machines:** Agents serve as intermediaries, enabling humans to interact with complex systems easily. They interpret human input and translate it into actionable tasks for machines. For example, a customer service chatbot helps users resolve issues by interacting with underlying databases and systems.

3.6.1.5 Properties of AI Agents

The properties of AI agents describe their abilities and how they behave, enabling them to work efficiently in changing and unpredictable environments. Designing and building these agents is a key part of creating intelligent systems used in many different fields. Some important properties are discussed below:

- **Actuation:** Agents utilize effectors or actuation mechanisms to execute actions based on their decisions, enabling them to influence or modify their environment.
- **Autonomy:** Agents operate independently, making decisions and taking actions without direct human involvement. They possess a level of independence in their decision-making processes.
- **Learning:** Many agents can learn from their experiences, either through explicit programming or by adapting based on feedback and interactions with their environment.
- **Perception:** Agents use sensors or perception mechanisms to observe and collect information about their environment. This information guides their decision-making and interaction with the surroundings.
- **Rationality:** Agents strive to make decisions that maximize expected utility or achieve the best possible outcome based on their knowledge and goals. Rational agents aim for optimal decision-making.
- **Goal-Directed Behavior:** Agents are designed to pursue specific objectives. Their actions and decisions are focused on achieving these goals, which can be predefined or learned over time.
- **Adaptability:** Agents adjust their behavior to respond effectively to changes in their environment or variations in task requirements. This often involves learning from experience.
- **Situational Awareness:** Agents maintain an understanding of their environment's current state, allowing them to adapt their behavior accordingly. This involves processing environmental information and assessing the context.
- **Communication:** Agents can communicate with other agents or external systems to exchange information, coordinate actions, or collaborate on tasks. This capability facilitates multi-agent systems and cooperative efforts.

- **Flexibility:** Agents demonstrate the ability to adapt to different situations, handling various tasks or modifying their strategies based on the complexity of the problem at hand.

3.6.2 AGENT ENVIRONMENT IN AI

An environment includes everything around an agent that is not part of the agent itself. It can be understood as the setting or situation where the agent exists and performs its actions.

The environment is where the agent operates, interacts, and gathers information through its sensors. It provides the agent with the context to sense and take actions. In most cases, the environment is considered **non-deterministic**, meaning its outcomes may not always be predictable or fully controlled by the agent.

Few examples of Agents and their Environments

Agent	Environment
Self-driving car	Roads, traffic, pedestrians, traffic signals, weather conditions
Virtual assistant (e.g., Alexa, Siri)	User inputs, device hardware, connected smart home devices, internet services
Chess-playing AI	Chessboard, opponent's moves
Robotic vacuum cleaner	Floors, furniture, walls, and obstacles in a house
Drone delivery system	Airspace, delivery route, weather, and obstacles (e.g., buildings, trees)
Email spam filter	Incoming emails, email metadata, and content of the messages
Surveillance camera system	Physical area under surveillance, lighting conditions, and movement detection

3.6.2.1 Characteristics of Agent Environment

An **environment** is everything an agent interacts with, including all external factors that affect its behavior. It can be defined by several characteristics:

1. **Fully Observable vs Partially Observable**
   - **Fully Observable:** An environment where the agent can sense the complete state of the environment at any given time. It simplifies decision-making as the agent doesn't need to maintain an internal state or track history.
     Example: A chessboard where all pieces and their positions are visible.
   - **Partially Observable:** An environment where the agent can only sense part of the environment's state, requiring it to rely on memory or inference to act.
     Example: A self-driving car with limited visibility due to fog.
   - **Unobservable:** If the agent lacks sensors or cannot perceive any part of the environment, it is called unobservable.

2. **Deterministic vs Stochastic**
   - **Deterministic:** The next state of the environment is completely determined by the current state and the agent's action.

     Example: Solving a Sudoku puzzle.
   - **Stochastic:** The next state of the environment involves randomness and is not solely dependent on the agent's action.

     Example: Playing dice games.
   - **Note:** If the environment is deterministic and fully observable, the agent doesn't face uncertainty in decision-making.
3. **Episodic vs Sequential**
   - **Episodic:** The agent's actions are divided into independent episodes, and it only needs to consider the current percept to make decisions.

     Example: Image classification tasks where each image is processed independently.
   - **Sequential:** The agent's actions are interdependent, requiring memory of past actions to decide future steps.

     Example: Playing chess, where each move influences the future game state.
4. **Single-Agent vs Multi-Agent**
   - **Single-Agent:** Only one agent operates in the environment.

     Example: A vacuum robot cleaning a room.
   - **Multi-Agent:** Multiple agents interact within the environment, potentially competing or collaborating.

     Example: Autonomous drones coordinating to deliver packages.
5. **Static vs Dynamic**
   - **Static:** The environment remains unchanged while the agent deliberates.

     Example: A crossword puzzle, where the state remains fixed.
   - **Dynamic:** The environment can change during the agent's decision-making process, requiring constant observation.

     Example: Driving a car in traffic, where conditions evolve in real-time.
6. **Discrete vs Continuous**
   - **Discrete:** The environment has a finite number of percepts and actions.

     Example: Chess, where there are a limited number of possible moves.
   - **Continuous:** The environment allows an infinite range of percepts and actions.

     Example: Controlling a robotic arm or a self-driving car navigating a highway.
7. **Known vs Unknown**
   - **Known:** The agent has complete knowledge of how the environment operates, including the outcomes of all possible actions.

Example: A robot in a factory preprogrammed with task details.

- **Unknown:** The agent must learn how the environment works to perform tasks effectively.

  Example: A robot exploring an unfamiliar building.

- **Note:** A known environment can be partially observable, and an unknown environment can be fully observable.

8. **Accessible vs Inaccessible**
   - **Accessible:** The agent can obtain complete and accurate information about the environment's state.

     Example: An empty room whose state can be measured by temperature or lighting.

   - **Inaccessible:** The agent cannot fully perceive or gather accurate information about the environment.

     Example: Understanding the dynamics of an earthquake or event occurring on Earth.

3.7 ETHICAL ISSUES AND FUTURE OF AI

The ethical issues surrounding Artificial Intelligence (AI) are vast and complex, as AI systems increasingly influence many aspects of human life. Here are some key ethical concerns and the future of AI:

3.7.1 ETHICAL ISSUES IN AI

- **Bias and Discrimination:** AI systems are often trained on data that reflects human biases, leading to biased decision-making. For instance, AI used in hiring or criminal justice systems may perpetuate racial or gender biases, impacting vulnerable groups negatively. Ensuring fairness and inclusivity in AI models is a major challenge.
- **Privacy Concerns:** AI often requires vast amounts of personal data to function effectively. This raises concerns about data privacy, surveillance, and how personal information is collected, stored, and used. There is a need for robust data protection frameworks to prevent misuse and breaches.
- **Autonomy and Accountability:** As AI systems are used in critical sectors such as healthcare, transportation, and finance, determining who is accountable for their actions becomes a significant issue. For instance, if an autonomous vehicle causes an accident, who is responsible? Legal frameworks need to address these questions.
- **Job Displacement:** The automation potential of AI could displace many jobs, particularly in industries such as manufacturing, customer service, and logistics. While AI creates new job opportunities, there is concern over how to retrain workers and manage economic disparities caused by this technological disruption.

- **Manipulation and Control:** AI technologies, such as deepfakes and personalized recommendation algorithms, can be used to manipulate individuals' opinions and behaviors. For example, AI-driven content creation can be used for political propaganda or spreading misinformation.
- **AI in Warfare:** The development of autonomous weapons and military AI systems raises ethical concerns about the potential for AI to make life-and-death decisions. The use of AI in warfare could lead to unforeseen consequences and ethical dilemmas related to accountability and the rules of engagement.
- **Existential Risks:** There is a philosophical concern about the long-term development of superintelligent AI systems. If AI becomes vastly more intelligent than humans, it could surpass human control, leading to scenarios where AI's goals conflict with human interests. Ensuring that AI remains aligned with human values is a crucial focus for AI researchers.

3.7.2 THE FUTURE OF AI

- **Human-AI Collaboration:** Rather than replacing humans, AI could augment human capabilities, leading to more productive and creative collaborations. AI could assist in areas such as healthcare, education, and research, enhancing human performance and decision-making.
- **Explainable AI:** One key future development will be the focus on making AI systems more interpretable and transparent. This would help users understand how AI makes decisions, which is vital for accountability and trust, especially in high-stakes areas like healthcare and criminal justice.
- **Regulation and Governance:** As AI becomes more pervasive, governments and international bodies will likely implement regulations to ensure its ethical use. These regulations may focus on transparency, data privacy, fairness, and accountability. We may see global agreements on AI ethics, similar to those already in place for nuclear technology.
- **AI for Good:** The future of AI holds significant potential for solving global challenges, such as climate change, healthcare, and poverty. AI could be used to optimize energy usage, assist in medical diagnosis, or help identify and address social issues. The "AI for Good" movement aims to harness AI's power to improve societal outcomes.
- **Ethical AI Design:** Future AI systems will likely incorporate more ethical decision-making frameworks, including systems that consider not only efficiency but also ethical principles, human values, and societal impacts. This approach will need input from diverse stakeholders, including ethicists, policymakers, and affected communities.
- **AI and Creativity:** As AI advances, it will likely continue to push boundaries in creative fields, including art, music, and literature. AI

a tool for creative expression, but it will raise questions about authorship, originality, and intellectual property rights.

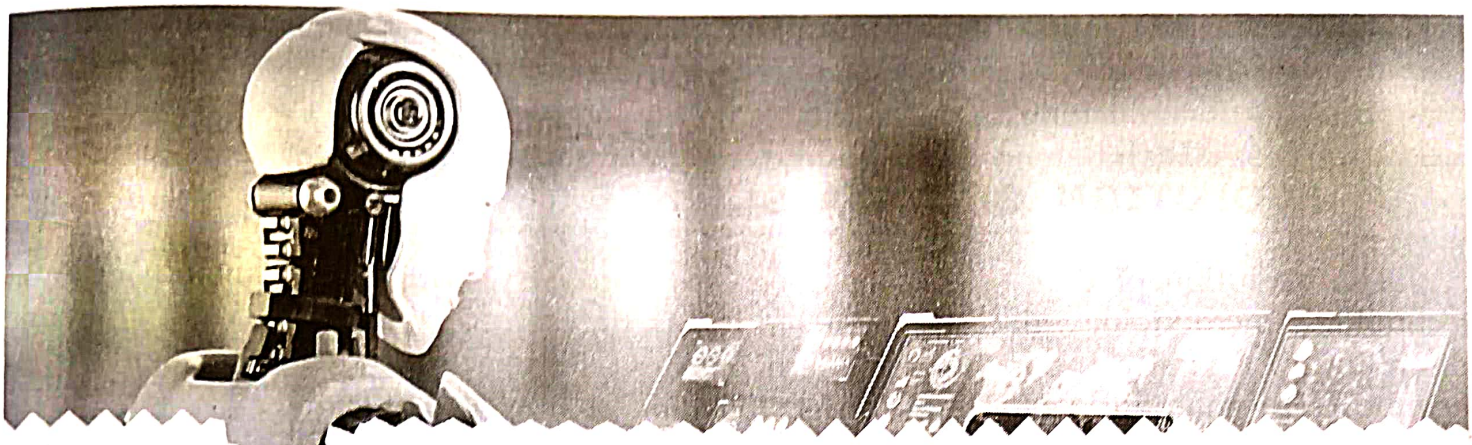
- **AI in Healthcare:** AI has immense potential to revolutionize healthcare by improving diagnostics, personalizing treatment plans, and even discovering new drugs. However, ethical concerns about patient consent, data privacy, and the impact on healthcare professionals need careful attention.

In summary, while AI holds great promise for enhancing human life and solving global problems, its development must be guided by ethical principles to prevent harm and ensure equitable benefits. The future of AI is likely to be shaped by a combination of innovation, regulation, and responsible usage.

QUESTIONS

1. Write aplication of AI in various fields.
2. What are the application of AI in Education field?
3. What are the application of AI in Healthcare?
4. What are the application of AI in Agriculture?
5. What is Problem-Solving? Explain various steps of problem solving using examples.
6. What do you mean by Agents? Explain the types of Agents.
7. What is the Agent Environment in AI? How does it influence the working of Agent?
8. Give a basic structure of AI agent and its characteristics.
9. Discuss the ethical issues and future of AI.

☆☆☆☆☆



CHAPTER

4 INTRODUCTION TO DATA SCIENCE

CHAPTER AT A GLANCE

4.1	Data Science	4.5	Data Science Related Fields
4.2	Evolution of Data Science	4.6	Applications of Data Science
4.3	Need of Data Science	4.7	Data Science Life Cycle
4.4	Types of Data	4.8	Careers in Data Science

4.1 DATA SCIENCE

Data science is an interdisciplinary field that uses statistical techniques, machine learning, and computational tools to extract meaningful insights from raw data. It plays a crucial role in solving complex real-world problems by enabling decision-making across industries such as healthcare, finance, e-commerce, and more. Data science has become essential in today's digital age, where vast amounts of data are generated daily across industries. Its ability to extract meaningful insights from raw data empowers organizations to make informed, data-driven decisions. By combining statistical analysis, machine learning, and advanced visualization, data science uncovers patterns and trends that traditional methods cannot. It plays a critical role in solving complex problems, optimizing processes, and driving innovation. From personalized healthcare to fraud detection in finance, data science enables tailored solutions that improve efficiency and outcomes. Moreover, as businesses strive to remain competitive, leveraging data science is no longer optional—it's a strategic necessity. Its versatility, coupled with its potential to address real-world challenges, makes data science one of the most transformative fields of the 21st century. At its core, data science involves collecting, cleaning, and analyzing large datasets to identify patterns, trends, and relationships. Techniques like predictive modeling, natural language processing, and data visualization help translate data into actionable knowledge. With the rapid growth of data generation and advanced tools, data science has become

a key driver of innovation. By bridging the gap between data and decision-making, it empowers organizations to optimize operations, improve customer experiences, and uncover new opportunities.

4.2 EVOLUTION OF DATA SCIENCE

The evolution of data science has been extraordinary, with significant advances in technology, methodology, and applications. It has progressed from a specialist academic discipline to a vital component of modern business, influencing industries ranging from healthcare to entertainment. Here's a detailed look at how it evolved:

Early Beginnings (pre-1960s) : Data science originated with statistical approaches developed in the 17th and 18th centuries. During this time, data was manually collected and recorded, mostly for government, trade, and scientific purposes. The history of data science may be traced back to the beginnings of statistics and mathematics. Data analysis was mostly done using basic arithmetic and statistical approaches including averages, standard deviations, and simple probability. The emphasis was mostly on descriptive statistics, which enabled analysts to summarise and present data in useful ways. Manual calculations, early statistical tables, and hand-drawn diagrams were among the fundamental tools available at the time. The concept of "data" was limited, frequently confined to small datasets and mostly academic or government research.

The Rise of Computational Tools (1960s–1980s) : In the 1960s and 1970s, the advanced computer technology, allowing data analysis operations to be automated. As computing power expanded, data storage technologies progressed from punch cards to early databases, and data processing software became more available. This era witnessed the creation of software tools such as SAS (Statistical Analysis System), which enabled more complex data analysis.

In the 1980s, the focus of statistics shifted from descriptive approaches to inferential statistics, which tried to make predictions or draw inferences about a population using sample data. Early statistical software was mostly employed by academics, government officials, and huge enterprises.

The Rise of Data Mining and Business Intelligence (1990s) : Data mining and business intelligence were introduced in the 1990s, bringing about a substantial transformation. Data mining techniques such as clustering, classification, and association rule mining have enabled firms to extract useful insights from vast datasets. During this time, the field of business intelligence evolved, which assists firms in making decisions based on data. This was the era when the first significant data warehouses were created, allowing businesses to store and query massive volumes of structured information.

Data analysts began querying relational databases using SQL (Structured Query Language), and the introduction of data visualization tools such as Tableau and Microsoft Excel enabled firms to interpret data more intuitively. However, the methodologies remained somewhat traditional, with an emphasis on organized data and batch processing.

The Development of Big Data and Machine Learning (2000s) : The explosion of data in numerous forms, including structured, unstructured, and semi-structured data, made the 2000s which is a transformative decade for data sci

internet, e-commerce, social media, and the Internet of Things (IoT) resulted in the generation of massive volumes of data, commonly referred to as "big data." Hadoop and MapReduce were created to handle, process, and analyze enormous datasets across distributed systems, ushering in a transition from traditional databases to big data frameworks.

Simultaneously, machine learning began to gain traction as a potent technique for forecasting future patterns using data. The combination of big data and machine learning enabled data scientists to create predictive models capable of analyzing and forecasting trends in real time. For example, e-commerce companies such as Amazon and Netflix used recommendation algorithms to propose purchases or movies to customers based on their previous activity.

During this time, the phrase "data science" gained popularity. It was recognized as a multidisciplinary domain that uses computer science, statistics, and domain experience to derive actionable insights from data. The combination of big data and machine learning enabled data scientists to create predictive models capable of analyzing and forecasting trends in real time.

During this time, the phrase "data science" gained popularity, reflecting the field's broader reach. It was recognized as a multidisciplinary domain that uses computer science, statistics, and domain experience to derive actionable insights from data.

Data Science as a Business Discipline (2010s–Present) : In the 2010s, data science matured as a business-critical discipline. The advent of cloud computing platforms such as Amazon Web Services (AWS), Google Cloud, and Microsoft Azure made data science more accessible to businesses of all kinds. Cloud-based solutions enabled the storage, processing, and analysis of massive volumes of data without requiring costly on-premise equipment.

As corporations recognized the usefulness of data, data science emerged as a key driver of commercial decision-making. Advanced analytics and machine learning models have been implemented by organizations across industries, including healthcare, banking, marketing, and entertainment, in order to optimize operations, tailor services, and improve consumer experiences.

Python, R, and Tensor Flow are examples of data science tools and languages that have become widely used, providing user-friendly environments for developing and deploying complex machine learning models. Data scientists and engineers began to collaborate more closely as new jobs emerged, such as data engineers, who focus on data pipelines and infrastructure, and machine learning engineers, who specialize in model deployment.

The future of data science is set to expand with advances in quantum computing, real-time data processing, and more automated machine learning systems. The growing integration of data science with other fields, such as the Internet of Things (IoT) and blockchain, will continue to open up new possibilities for boosting efficiency, generating growth, and addressing global concerns.

4.3 NEED OF DATA SCIENCE

Data Science has become an indispensable field in today's data-driven world, where massive amounts of information are generated every second. It addresses the need to extract valuable insights from this data, helping individuals, businesses, and governments make informed decisions. Here's why Data Science is crucial:

1. **Big data Era :** The world generates an overwhelming volume of data every second from sources like social media, IoT devices, business transactions, and sensors. This "big data" is characterized by its high volume, velocity, and variety, making traditional methods insufficient. Data science provides tools and frameworks to manage and analyze these vast datasets efficiently.
2. **Complexity of Data :** Data comes in various forms, such as structured (spreadsheets), unstructured (text, images, videos), and semi-structured (JSON, XML). Handling this diversity requires specialized techniques, such as natural language processing (NLP) for text and computer vision for images. Data science bridges the gap between raw data and actionable insights.
3. **Decision Making :** Organizations must base their decisions on evidence rather than intuition. Data science uses historical data to identify patterns, correlations, and trends, enabling businesses to forecast outcomes and make informed choices in areas like marketing, inventory management, and risk assessment. Companies like Netflix, Amazon, and Spotify use Data Science to create personalized recommendations, enhancing customer satisfaction. Tailoring experiences based on user behavior and preferences improves engagement and loyalty.
4. **Improved Efficiency :** Data science automates repetitive tasks, streamlines operations, and optimizes resource allocation. For example, in logistics, data science helps determine the most efficient delivery routes, reducing costs and time. In manufacturing, it identifies bottlenecks to improve productivity. Data Science powers intelligent systems such as chatbots, virtual assistants, and recommendation engines. Automation saves time and resources by streamlining repetitive tasks while providing high accuracy.
5. **Predictive and Prescriptive Analytics :** Predictive analytics forecasts future events based on historical data, such as sales trends or stock price movements. Prescriptive analytics goes a step further by recommending actions to achieve desired outcomes, such as optimal pricing strategies or risk mitigation plans.
6. **Real-Time Problem Solving :** Many industries require instantaneous responses to problems, such as fraud detection in financial transactions or monitoring network security. Data science enables real-time analysis and decision-making, ensuring quick resolution and minimal disruption.
7. **Innovation:** Data science drives innovation by analyzing user behavior and feedback, leading to the development of new products and services. It also enables personalization, such as Netflix recommending shows based on viewing history or e-commerce platforms suggesting products tailored to individual preferences.

8. **Competitive Advantage :** In today's competitive landscape, data-driven strategies are a key differentiator. Companies using data science can anticipate market trends, understand consumer preferences, and tailor products or services to meet demands, giving them a significant edge over competitors.
9. **Cross-Industry Applications:** Data science is not limited to one domain. It is used in healthcare for early disease detection, in finance for credit scoring and portfolio management, in education for adaptive learning systems, and in environmental science for climate modeling and resource management.
10. **Enabling Ethical and Sustainable Growth :** Data Science supports sustainability initiatives by analyzing environmental impact data and optimizing resource use. It also helps in addressing social issues, such as identifying and reducing inequality or improving public services.

4.4 TYPES OF DATA

Data is the foundation of data science and is utilized for decision-making and extract meaningful insights. Understanding the different types of data is critical in data science because it allows you to choose the best approaches and tools for analysis, modeling, and visualization. Data can be divided into five categories: structured, unstructured, semi-structured, quantitative, and qualitative.

Structured Data: Structured data is data that is structured in a specified format, such as tables with rows and columns. It is well-organized and simple to enter, store, query, and analyze, making it the most straightforward sort of data to process. Structured data is typically kept in relational databases or spreadsheets and follows a predefined schema, with each data piece assigned to a distinct field. Following are some examples of structured data :

- **Customer Data :** Customer data in a Customer Relationship Management (CRM) system includes fields such as customer ID, name, address, phone number, email, and purchase history, which are arranged into database tables.
- **Sales Transaction Data :** Structured data in retail provides transaction details such as product ID, quantity, price, purchase date, payment method, and customer information.
- **Financial Data :** Financial data, such as balance sheets, income statements, and transaction records in accounting systems, are structured in rows and columns, with each row representing a single transaction or record and each column representing a specific attribute.
- **Employee Records:** Structured data about employees, such as their ID, name, department, job title, salary, hire date, and performance ratings, is often recorded in HR management systems.
- **Inventory Management :** Inventory management is the organization of data on products in a warehouse into a relational database, including product ID, description, quantity in stock, reorder level, and supplier information.

- **Healthcare Data:** Patient data in electronic health record (EHR) systems include fields such as patient ID, name, age, medical history, diagnosis, and therapy, which are neatly grouped into structured tables.

Unstructured Data: Unstructured data is more complex and does not have a set format or organization. This type of data is not arranged in a way that allows for immediate search and analysis. It include text, photos, music, video, social media posts, and site pages. Unstructured data analysis frequently necessitates the use of advanced techniques such as natural language processing (NLP) for text analysis and computer vision for image and video processing. While unstructured data accounts for the majority of data collected today, it is often more difficult to process and requires specialized algorithms to yield valuable insights. Following are some examples of unstructured data :

- **Text Data:** This comprises information from blogs, forums, social media posts, and emails. The lack of a predetermined format makes it challenging to assess the content directly without additional processing. To analyze this data, natural language processing (NLP) techniques are frequently employed.
- **Images:** Unstructured data includes pictures, drawings, and other image files (such as JPEG, PNG, GIF, etc.). In order to get valuable insights, such as object detection or facial recognition, from the analysis of these files, methods like picture recognition or computer vision algorithms are frequently needed.
- **Audio Files:** Audio data is unstructured and includes things like music, podcasts, and recorded conversations (MP3, WAV, etc.). In order to facilitate further analysis, speech recognition technologies are frequently employed to translate spoken words into text.
- **Video Content:** Both visual and aural information can be found in videos, including surveillance footage, YouTube videos, and video conferences. It is frequently necessary to examine the content using both voice and picture recognition techniques in order to extract meaningful information from these files.
- **Web Pages:** Text, pictures, videos, and links are all examples of unstructured information on web pages. A webpage's overall structure (HTML) is not structured in a way that makes it easy to study without processing, even though web scraping tools can extract certain data.
- **Customer Reviews:** Examples of unstructured data include product reviews, free-text survey responses, and reviews posted by customers on e-commerce websites. To find insights, they call for text mining or sentiment analysis methods.

Semi-structured Data : Semi-structured data falls between structured and unstructured data. While it may not fit neatly into the rows and columns of a standard database, it does have some organizational components such as tags or markers to distinguish data fields. This type of data has some structure but is adaptable enough to store a wide range of information. Semi-structured data includes JSON files, XML files, and email messages. Data fields are commonly designated by tags (such <name> or <email>), although the structure might vary between instances. Following are some examples of semi-structured data :

- **JSON (Javascript Object Notation) Files** : JavaScript Object Notation, or JSON, files are frequently used to transfer data between a web application and a server. Although they don't adhere to relational databases' rigid tabular structure, they do have tags or keys (such "name," "age," or "address") that facilitate data organization and interpretation.
- **XML (eXtensible Markup Language) Files:** Data can be stored and transferred in an organized yet adaptable fashion using XML (eXtensible Markup Language) files. It makes data elements understandable by both humans and machines by defining them with tags. For example product ID, price, and description may be stored in XML format by an e-commerce platform; however, each product may have unique properties.
- **Email:** One typical instance of semi-structured data is an email. Emails have specified fields for sender, recipient, date, and topic, but the body of the email is unstructured and can contain links, graphics, or plain text. The body content can vary greatly, but the headers give it structure.
- **No SQL Databases:** Semi-structured data is stored in databases like CouchDB and MongoDB. No SQL databases enable data to be kept in flexible formats, such as documents or key-value pairs, in contrast to relational databases, which store data in tables with rigid schemas. Even if each record may have a different set of fields or attributes, they can still be quickly queried and processed.
- **Log Files:** Both organized and unstructured data are frequently found in system or application logs. For example, a server log may include unstructured text descriptions of events or faults in addition to structured timestamps, IP addresses, and error codes, resulting in semi-structured data.
- **CSV Files with Inconsistent Data:** Although CSV (Comma-Separated Values) files may seem structured at first, they may contain optional fields or inconsistent columns, particularly when the data source is not standardized. Because it is somewhat organized yet might change between rows or entries, it is considered semi-structured.

Quantitative Data : Quantitative data includes numerical values that can be measured and quantified. It is widely utilized in statistical analysis, predictive modeling, and decision-making. Quantitative data can be classified as discrete or continuous. Discrete data relates to countable amounts (for example, the number of products sold), whereas continuous data represents measurable quantities that can take any value within a given range. This information is critical for developing insights that require precise measurements, such as financial predictions or scientific investigations. Some of the examples of quantitative data are as under :

- **Age:** Age can be measured and expressed as a numerical number, an individual's age, such as 20 or 40 years old, is considered quantitative data.
- **Income:** Quantitative information is provided on a person's or household's income, such as ₹ 50,000 per month. It can be utilized for statistical analysis and is quantifiable.

- **Height and Weight:** They can be measured and quantified, height (for example 170 cm) and weight (for example 70 kg) are instances of quantitative data.
- **Temperature:** Quantitative information includes temperature readings such as 22°C or 98.6°F. They are a quantifiable value that can be applied to a variety of analysis tasks, including medical measurements and climatic studies.
- **Distance or Length:** Distance is a quantitative variable that may be measured with a standard unit, such 200 meters or 10 kilometers.
- **Number of Products Sold:** Quantitative data is the quantity of units sold, such as 500 smartphones or 1000 books.
- **Population Size:** The population of a city or nation, such as three million, is quantitative data.
- **Speed:** Speed it is a quantifiable rate of movement, like 40 kilometers per hour, is quantitative.

Qualitative Data : Qualitative data, often known as categorical data, is a collection of non-numerical values used to identify and label variables. It is often descriptive in nature, aiding in the classification of data into various groups.

Qualitative data can be further classified into nominal and ordinal categories. Nominal data lacks an intrinsic order or ranking (e.g., colors, countries, or product types), whereas ordinal data has a specified order but no fixed intervals between values. While qualitative data is not appropriate for mathematical operations, it is essential for grouping and evaluating trends in domains such as market research, social sciences, and customer experience. Following are some examples of qualitative data:

- **Customer feedback:** Unstructured answers provided by clients in questionnaires or evaluations are qualitative. A consumer may comment, for example, "The product quality is excellent" or "The delivery was too slow." These are qualitative because they are written descriptions that express thoughts, emotions, or experiences.
- **Survey Responses:** Likert scales and categorical questions are frequently used in surveys to collect qualitative data. If someone were asked, "What is your favorite type of music?" for instance, they would say, "Rock, " "Jazz, " or "Classical." These non-numerical categories reflect the traits or inclinations of the participants.
- **Interview Transcripts:** Data gathered from interviews in which participants give detailed responses to open-ended questions. For instance, qualitative information like "I am very passionate about technology, " which is descriptive and difficult to quantify, may be revealed during a job interview.
- **Product Types and Categories:** Information on the kinds of products, like "Electronics, " "Clothing, " or "Furniture." These are categorical variables without numerical meaning that are used to group products according to their characteristics.

- **Social Media Posts :** Posts or tweets on social media sites may include words like "happy," "excited," or "disappointed," which are qualitative indicators of feelings or sentiments and are usually subjected to sentiment analysis.

4.5 DATA SCIENCE RELATED FIELDS

Data science is an interdisciplinary field that brings integrates concepts, tools, and approaches from several fields to extract knowledge and insights from data. Its adaptability means that it overlaps with and supports a variety of related fields, each with a distinct specialty.

4.5.1 ARTIFICIAL INTELLIGENCE

Data science and artificial intelligence (AI) are closely related fields, with data science serving as the foundation for developing intelligent systems. AI focuses on developing machines with human-like decision-making and problem-solving abilities, whereas data science provides the data, tools, and methodologies required to train these systems. Data science, which employs approaches such as statistical analysis, machine learning, and natural language processing, assists AI models in learning from large datasets to spot patterns, make predictions, and automate operations. The collaboration of data science and AI converts data into intelligent solutions that drive innovation across industries.

4.5.2 STATISTICS

Statistics is the foundation of data science, offering the mathematical framework and tools required to evaluate and understand data efficiently. Statistics plays an important role in data science by assisting with experiment design, summarizing enormous datasets, spotting patterns, and generating predictions. Hypothesis testing, regression analysis, and probability modeling are critical methods for confirming insights and developing accurate models. Statistics helps data scientists to draw meaningful conclusions from noisy or incomplete data by quantifying uncertainty and providing occuracy in analysis. This collaboration between statistics and data science enables organizations and researchers to confidently make data-driven decisions, connecting raw data to actionable insights.

4.5.3 MACHINE LEARNING

Machine learning (ML) is a critical component of data science, allowing computers to learn from data and make predictions or judgments without requiring explicit programming. In the data science pipeline, machine learning algorithms play an important role in automating pattern recognition, classification, regression, clustering, and anomaly detection. Data science uses machine learning models such as decision trees, neural networks, and support vector machines to tackle difficult problems in industries such as healthcare, banking, and ecommerce. Machine learning transforms raw data into meaningful insights through iterative data preprocessing, feature selection, and model tuning, enabling predictive and prescriptive analytics that promote industry-wide innovation and efficiency.

4.5.4 DATA MINING

Data mining is an important component of data science that focuses on identifying patterns, correlations, and insights from massive databases. Data mining is a fundamental approach that allows data scientists to explore and analyze organized and unstructured data using techniques like clustering, classification, association rule learning, and anomaly detection. It acts as a link between raw data and actionable knowledge, assisting businesses in identifying trends, optimizing tactics, and forecasting outcomes. Data mining, when integrated into the broader data science lifecycle, enables applications like as customer segmentation, market basket research, and fraud detection by translating complex datasets into important decision-making insights.

4.5.5 BIG DATA

Data science and big data are closely related. Data science offering the approaches and tools required for extracting insights from enormous, complicated information. Big data refers to massive amounts of data generated at rapid speeds from many sources such as social media, IoT devices, and enterprise systems. Data science efficiently processes and analyzes this data using advanced analytics, machine learning, and distributed computing frameworks such as Hadoop and Spark. Data science transforms raw information into actionable insights by solving big data's concerns of volume, variety, velocity, validity, and value, allowing businesses to make data-driven decisions, optimize operations, and discover new possibilities in an increasingly data-centric world.

4.5.6 BUSINESS INTELLIGENCE

Data science and business intelligence work hand-in-hand to enable organizations to make data-driven decisions. While Business Intelligence focuses on summarizing historical data through dashboards, reports, and visualization tools, data science takes it further by employing predictive and prescriptive analytics to forecast trends and optimize outcomes. Data science enriches business intelligence by applying advanced techniques like machine learning, statistical modeling, and natural language processing to uncover hidden patterns, relationships, and opportunities in data. This collaboration transforms static reports into dynamic insights, empowering businesses to anticipate customer needs, enhance operational efficiency, and drive strategic innovation in a competitive marketplace.

4.5.7 NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a critical subfield of data science, focusing on enabling machines to understand, interpret, and generate human language. By combining linguistics, machine learning, and statistical techniques, data science provides the tools to process and analyze textual data effectively. Tasks such as sentiment analysis, text classification, language translation, and chatbot development rely on NLP to extract meaningful insights from unstructured text data. Data science plays a pivotal role in preprocessing text, training models, and validating results, making NLP applications possible in areas like customer service, content analysis, and voice recognition. This collaboration enhances human-computer interaction and unlocks the potential of language in driving innovation.

4.5.8 CLOUD COMPUTING

Cloud computing has transformed data science by delivering scalable, adaptable, and cost-effective resources for storing, processing, and analysing massive datasets. Data science makes use of cloud platforms such as AWS, Google Cloud, and Microsoft Azure to gain access to powerful computational tools and frameworks that would otherwise need costly on-premise infrastructure. These platforms facilitate the seamless integration of data pipelines, machine learning model deployment, and real-time analytics. Cloud computing also makes collaboration easier by enabling distant teams to work on common datasets and models. This collaboration enables data scientists to effectively handle massive data, deploy AI solutions at scale, and deliver actionable insights, transforming enterprises in the digital age.

4.5.9 INTERNET OF THINGS (IOT)

Data science and IoT (Internet of Things) work together to unlock the full potential of interconnected devices, sensors, and systems. IoT generates massive volumes of data in real-time from diverse sources like smart homes, wearable devices, and industrial machinery. Data science plays a critical role in processing, analyzing, and deriving actionable insights from this data. Techniques such as predictive analytics, anomaly detection, and pattern recognition help optimize operations, enhance user experiences, and improve decision-making. Various applications include predictive maintenance in manufacturing, energy consumption optimization in smart grids, and health monitoring in wearable devices. This integration drives innovation, enabling smarter and more efficient IoT ecosystems.

4.5.10 COMPUTER VISION

Data science is critical in the subject of computer vision, which allows machines to interpret and understand visual data from the real world, such as photographs and videos. Computer vision systems can detect objects, recognize faces, analyze sceneries, and even do picture segmentation and optical character recognition (OCR) using data science approaches such as machine learning, deep learning, and image processing. Data science is critical for developing computer vision models, optimizing algorithms, and evaluating massive amounts of visual data. This integration has applications in a variety of industries, including autonomous vehicles, medical imaging, security monitoring, and retail, advancing automation and improving human-computer interactions.

4.6 APPLICATIONS OF DATA SCIENCE

Data Science involves huge quantity of data used in deep analysis. The importance of data science exists in almost every field nowadays. The applications of data science in various fields are as under:

4.6.1 SEARCH ENGINES

Today search engines are the backbone of Internet system. Everything on Internet nowadays can be accesses through search engines. In the race of fast processing, every

search engine uses new techniques time to time to reduce the searching time. Data Science mainly used in search engines for the fast processing of results. Without data science, it is nearly impossible to handle large amount of data. Data science ensures that search engines continually evolve, offering faster, smarter, and more personalized experiences for users worldwide. Using data science algorithms, search engines offer the best results in a matter of seconds. Search engines like Google, Bing, and Yahoo leverage data science extensively to provide accurate, relevant, and personalized results. The key applications of data science in search engines include:

- **Ranking Algorithms:** Data science algorithms like PageRank and BERT to rank web pages based on relevance, user intent, and content quality.
- **Query Understanding :** Natural Language Processing (NLP) techniques interpret user queries, accounting for synonyms, context, and ambiguity to improve search precision.
- **Personalized Recommendations:** Machine learning analyzes user history and preferences to deliver tailored search results, advertisements, and content recommendations.
- **Autocomplete and Suggestions :** Predictive modeling anticipates user queries and provides auto-complete suggestionses to enhance search efficiency.
- **Advertisement Targeting :** Data science identifies user intent and demographics to display targeted advertisements to maximize advertisement relevance and revenue.
- **Handling Multilingual Searches :** Advanced NLP and translation models enable effective searches in multiple languages to break linguistic barriers.
- **Image and Video Search :** Computer vision techniques process and index visual content allow users to search using images or videos.
- **Spam Detection and Filtering :** Data science helps identify and filter out low-quality or malicious content to ensure the reliability of search results.
- **Voice Search Optimization :** Voice recognition and NLP technologies enable seamless voice-based searches to cater to diverse user needs.

4.6.2 RECOMMENDATION SYSTEMS

In the era of Internet, the availability of infinite online content can become a big problem. Everyone wants the contents as per their choice. To search contents as per the need, the process is cumbersome and time-consuming. Data Science helps the users to meet this gap by suggesting the only contents that the user might be interested. For example, various online movies platforms suggest movies based on what a user like to watch. E-Commerce websites suggest products recommendation based on what a user like to purchase. In E-Commerce and entertainment websites, data science algorithms work on user behaviour about what a user is searches for, what a user is interested in, what links they clicked in or what movies they watch. Recommendation systems are a key application of data science, providing personalized suggestions that enhance user experience across various industries.

The key applications of data science in Recommendation systems include:

- **E-commerce :** Data science personalized product recommendations based on user interests, past browsing activity, and purchase behavior. It recommends related or improved products in order to upsell and cross-sell.
- **Streaming Platforms :** It provides content recommendation for movies, TV shows, or music using user preferences and behavior analysis. It suggests dynamic playlists based on individual or collective user patterns.
- **Social Media :** It suggests friends, connections, or groups to users. It recommend posts, videos, or trending topics based on user interests.
- **Online Learning Platforms :** It provides tailoring course recommendations based on previous enrollments, learning pace, and interests. It suggests resources like articles, quizzes, or tutorials for skill enhancement.
- **Retail and Grocery Stores :** It suggests products during online shopping or mobile app browsing. It personalize discounts or coupons for frequently purchased items.
- **Healthcare** : It recommends personalized wellness plans, exercises, or medication reminders based on patient history. It suggests medical articles, resources, or specialists.
- **Tourism and Travel** : It provides tailoring travel destinations, hotel options, or activities based on user preferences and past bookings. It recommends travel packages or local attractions during itinerary planning.
- **Job Portals** : It suggests job openings based on user skills, preferences, and application history. It Offers career-related insights and training courses.

4.6.3 DELIVERY LOGISTICS

Logistic companies use data science to optimize routes through which the product can be delivered on time. It also suggests the best time to deliver the item and best mode of transportation. ata science has revolutionized delivery logistics by optimizing operations, enhancing efficiency, and improving customer experiences. The key applications of data science in delivery logistics include:

- **Route Optimization** : Algorithms analyze real-time traffic, weather conditions, and delivery constraints to identify the most efficient routes, saving time and fuel costs.
- **Demand Forecasting** : Predictive analytics helps anticipate order volumes, enabling better planning of resources like vehicles, drivers, and warehouse stock.
- **Last-Mile Delivery** : Data-driven models optimize last-mile delivery by clustering deliveries, selecting appropriate vehicles, and minimizing delays.
- **Warehouse Management** : Analyzing inventory data ensures accurate stock levels, reducing overstocking or shortages. Robotics and AI-powered systems streamline picking and packing processes.
- **Real-Time Tracking** : Data science enables dynamic tracking of shipments, providing real-time updates to customers and logistics teams.

- **Predictive Maintenance** : Machine learning monitors vehicle performance data to predict and prevent potential breakdowns, ensuring smooth operations.
- **Cost Optimization** : Analyzing delivery costs per route or order helps companies to minimize expenses and maximize profitability.
- **Customer Experience** : Personalized delivery options, time-slot predictions, and accurate eshmate time to travel are powered by data science, enhancing customer satisfaction.
- **Fraud Detection** : Data models identify anomalies in delivery patterns, preventing theft or loss in the supply chain.
- **Sustainability Initiatives** : Optimizing routes and delivery schedules reduces carbon emissions, supporting greener logistics practices.

4.6.4 FRAUD DETECTION

Financial industries always have an issue of fraud. To detect fraudulent transactions, these financial industries use data science algorithms. Professionals from computer networks security can use data science algorithms to protect system from security threats by looking for suspicious patterns. Data science plays a pivotal role in fraud detection by leveraging advanced algorithms and statistical techniques to identify and prevent fraudulent activities in real-time. Here are the key applications:

- **Anomaly Detection** : Machine learning models analyze patterns in transactions or user behavior to detect deviations from normal activity, such as unusual spending or login locations.
- **Predictive Modeling** : Historical data is used to train predictive models that can flag transactions likely to be fraudulent, enabling proactive measures.
- **Real-Time Monitoring** : Data science enables systems to analyze thousands of transactions in real-time, identifying suspicious activities as they occur.
- **Natural Language Processing (NLP)** : NLP is used to detect fraudulent claims or emails by analyzing the language patterns and inconsistencies in communication.
- **Network Analysis** : Graph-based techniques uncover relationships between entities, such as fraudulent accounts working together, to expose organized fraud networks.
- **Behavioral Analytics** : User behavior is continuously tracked to identify anomalies, such as multiple failed login attempts or unusual access times.
- **Risk Scoring** : Data science assigns risk scores to transactions or entities, helping organizations prioritize investigations and minimize false positives.
- **Credit Card Fraud** : Identifying unusual purchase patterns or locations that deviate from a cardholder's typical behavior.
- **Insurance Fraud** : Analyzing claims data to detect fabricated or exaggerated claims through pattern recognition.

- **E-commerce and Cybersecurity** : Preventing account takeovers, fake reviews, and payment fraud by analyzing user interactions and transaction logs.

By automating and enhancing fraud detection, data science reduces financial losses, protects users, and improves trust in digital systems. Its ability to adapt and evolve with new fraud techniques makes it indispensable in combating financial crime.

4.6.5 BANKING AND FINANCE

Banking and finance industries use data science algorithms to reduce risk of losses. Data science analytical tools are used by these financial institutions to predict the future. Data science is transforming the banking and finance industry by enabling smarter decision-making, optimizing processes, and enhancing customer experiences. Key applications of data science in banking and finance are as under:

- **Fraud Detection** : Advanced algorithms identify suspicious transactions, unusual account activities, and potential fraud in real-time, minimizing financial losses.
- **Risk Management** : Predictive models assess creditworthiness and loan repayment likelihood by analyzing historical data and customer profiles, reducing default risks.
- **Customer Analytics** : Data science helps banks understand customer behavior, enabling personalized services, targeted marketing, and enhanced customer retention.
- **Algorithmic Trading** : Machine learning models analyze market trends, financial data, to execute high-frequency trades with precision and speed.
- **Portfolio Management** : Data-driven tools optimize investment portfolios by assessing risks, returns, and market conditions, aiding both retail and institutional investors.
- **Loan and Credit Scoring** : Automated systems evaluate applicants' credit worthiness using a combination of financial history, spending habits, and behavioral data.
- **Regulatory Compliance** : Data science ensures compliance with financial regulations by automating reporting, identifying anomalies, and monitoring transactions for suspicious activity.
- **Chatbots and Customer Support** : AI-powered virtual assistants streamline customer interactions by resolving queries, providing financial advice and assisting with transactions.
- **Operational Efficiency** : Process automation through data science reduces manual tasks in areas like reconciliation, reporting, and underwriting, saving time and costs.
- **Fraud Prediction in Insurance** : Analyzing claims data to detect fraudulent insurance applications or exaggerated claims, improving the integrity of the system.

4.6.6 AUTOCOMPLETE

AutoComplete feature is widely used nowadays. The user gets the facility of autocomplete by just type few letters or words, and he will get the full sentence

automatically. Autocomplete is an important part of data science. For example, in Google mail, autocomplete feature suggests to complete the whole line just by typing few words by the user. Autocomplete, a feature found in search engines, text input fields, and messaging apps, uses data science to predict and complete a user's query or input before it is fully typed. It enhances user experience by speeding up interactions and improving accuracy. Key applications of data science in autocomplete include:

- **Predictive Text Modeling** : Natural Language Processing (NLP) techniques analyze vast amounts of text data (such as search history, common phrases, and language models) to predict the most likely word or phrase a user intends to type.
- **Contextual Suggestions** : Data science uses contextual analysis to provide autocomplete suggestions based on the current sentence or the user's prior inputs. For example, if a user starts typing "weather, " the system may suggest "forecast, "today" or "tomorrow" depending on the context.
- **Personalization** : By analyzing individual user behavior and historical data, autocomplete can provide personalized suggestions, making interactions more efficient. For instance, frequent searches or past queries influence future suggestions.
- **Spelling and Grammar Correction** : Data science algorithms use models trained on large text corpora to suggest the correct spelling or grammar as the user types, improving the accuracy of inputs.
- **Search Engine Optimization** : Autocomplete in search engines predicts queries based on trending topics, location, and user history. Machine learning models identify common search patterns and adjust suggestions to align with popular or frequently searched terms.
- **Real-Time Data Analysis** : Autocomplete systems process real-time data from users' input to provide immediate and relevant suggestions, using machine learning to adjust dynamically.
- **Multilingual Support** : Autocomplete systems are enhanced to support multiple languages, using data science for language detection, translation, and providing language-specific suggestions based on user inputs.
- **Query Understanding and Intent Prediction** : Advanced data models analyze a user's intent from partial input and suggest full queries, significantly improving search and navigation in applications. For example, typing "best restaurants" might suggest "best restaurants near me" or "best restaurants in [location]."
- **Voice-to-Text Autocomplete** : In voice-enabled applications, data science helps transcribe speech to text and provides autocomplete suggestions based on spoken input, adjusting for pronunciation and accent variations.

4.6.7 VIRTUAL ASSISTANTS

Virtual Assistants are very popular in these days. Data science algorithms plays an important role in virtual assistants. Virtual assistants use deep neural networks with natural language processing to hear, understand and execute the command verbally and

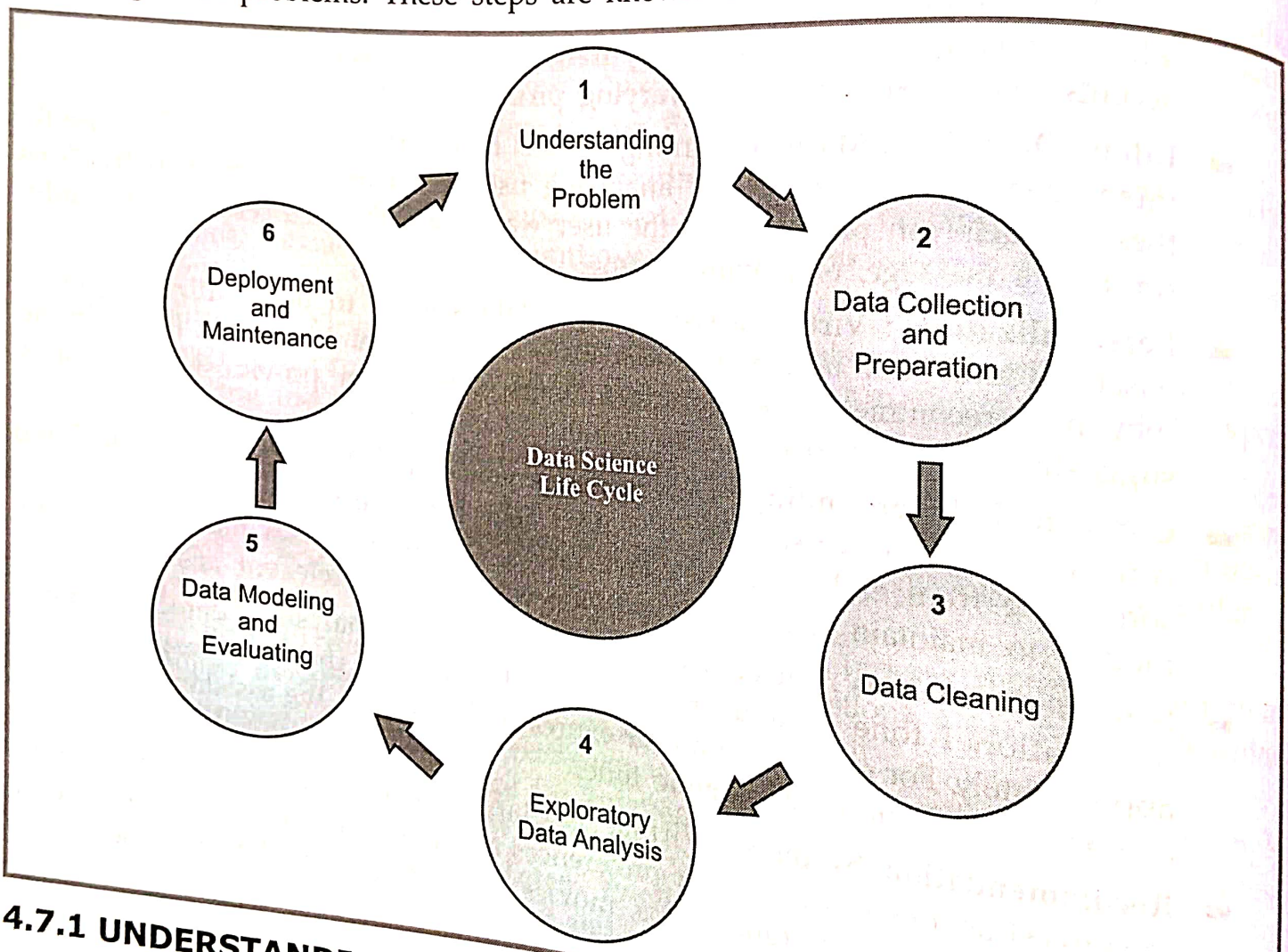
respond to the user verbally. Virtual assistants, such as Siri, Alexa, Google Assistant, and Cortana, have become an integral part of daily life, and data science plays a crucial role in enhancing their performance and effectiveness. Key applications of data science in virtual assistants include:

- **Natural Language Processing (NLP)** : Virtual assistants rely heavily on NLP to understand and process human language. Data science techniques, such as tokenization, named entity recognition, and syntactic parsings enable virtual assistants to interpret and respond to spoken or written queries accurately.
- **Speech Recognition and Voice-to-Text** : Data science algorithms convert voice commands into text. Machine learning models trained on large datasets of audio allow virtual assistants to accurately transcribe and understand speech, even with accents, background noise, or varying pronunciations.
- **Intent Detection** : Machine learning models help virtual assistants determine the intent behind user queries. By analyzing user input and historical interactions, these models can predict what the user wants, whether it's setting a reminder, sending a message, or playing a song.
- **Personalization** : Virtual assistants use data science to personalize responses based on individual user preferences, history, location, and behavior. For example, they may recommend specific songs, adjust settings, or provide location-based suggestions tailored to the user's habits.
- **Contextual Understanding** : Data science enables virtual assistants to understand context within conversations. If a user asks follow-up questions or refers to something from earlier in the conversation, the system uses machine learning models to maintain context and provide coherent and relevant answers.
- **Sentiment Analysis** : Virtual assistants use sentiment analysis to detect the mood or emotional tone of a user's input and allowing them to respond more appropriately. For example, if a user expresses frustration, the assistant may adopt a more empathetic or apologetic tone.
- **Recommendation Systems** : Virtual assistants use recommendation algorithms to suggest actions based on user preferences and past behavior. For instance, they may recommend restaurants, apps, movies, or music, tailored to the user's past choices and feedback.
- **Task Automation** : Data science helps virtual assistants to automate tasks like scheduling, sending messages, controlling smart home devices, and providing reminders. Virtual assistants predict tasks the user might want to complete and offer suggestions by analyzing patterns in user behavior.
- **Multi-Modal Capabilities** : Virtual assistants can combine voice recognition with other inputs, such as visual data. Data science algorithms process inputs from various sources (e.g., voice, text, images) to deliver seamless interactions across different modes. For example, assistants may recognize a product image and offer related information.

- **Continuous Learning and Improvement :** Data science enables virtual assistants to continually improve by learning from each interaction. Using machine learning models, they refine their understanding and responses over time and becoming more accurate and effective as they process more data.

4.7 DATA SCIENCE LIFE CYCLE

To solve a problem for an organization that contains large amount of data is not an easy task. Some sequential steps are followed by data scientists to complete the process of solving these problems. These steps are known as data science life cycle.



4.7.1 UNDERSTANDING THE PROBLEM

The first step of data science life cycle is to identify the business problem and to understand the problem with clear objective around with all the other steps will be articulated. Identifying the problem is one of the major step of data science life cycle. It involves understanding and framing the problem that needs to be solved in a clear and structured way. This step lays the foundation for the entire project, ensuring that efforts are aligned with the business or research goals and that resources are effectively utilized. This stage involves collaborating with stakeholders to understand the context, goals, and constraints of the problem, ensuring alignment with business or research

includes defining measurable success metrics, identifying the target variable and relevant predictors, and determining whether the problem involves descriptive, diagnostic, predictive, or prescriptive analysis. A well-defined problem considers the scope, assumptions, and feasibility, setting a clear direction for data collection, analysis, and modeling. This step lays the foundation for the project, ensuring efficient resource use, measurable outcomes, and solutions that address the true needs of stakeholders.

4.7.2 DATA COLLECTION AND PREPARATION

Once the problem is identified, the next step of data science life cycle is to collect the data. Data acquisition systematically collects data from various sources. Data can be collected from relevant sources. These include databases, APIs, Web scraping and public datasets. Data can be collected in structured or unstructured form. Usually, there are multiple steps in the data collecting process, all of which are intended to guarantee that the information gathered is correct, pertinent, and prepared for analysis. Data collection and preparation are critical processes in the data science life cycle, involving the gathering, cleaning, and arranging of data before analysis. It is critical to ensure that the data obtained is relevant, accurate, and indicative of the problem under consideration.

Once collected, the data must be prepared for analysis, which entails numerous critical processes. First, data is cleaned to manage missing values, errors, and duplicates. This ensures the dataset's accuracy and consistency. Next, the data is converted, which may include normalizing or scaling numerical values, encoding categorical variables, and generating new features from existing ones to improve the model's prediction ability.

Data preparation also involves data integration, which is the process of merging or joining various datasets from diverse sources to create a single dataset. This phase is critical to ensuring that the data is complete and suitable for exploratory analysis or modeling. Proper data collection and preparation are crucial to the success of a data science project because they result in a clean, structured dataset that allows for accurate, meaningful insights and trustworthy predictive models.

4.7.3 DATA CLEANING

Data cleaning is an important phase in the data science life cycle since it ensures that the dataset is correct, consistent, and ready for analysis. During this phase, data scientists work to detect and correct issues such as missing numbers, duplicates, errors, inconsistencies, and outliers that could violate the data's integrity. These problems are frequently caused by a variety of factors, such as data entry errors, sensor errors, or integration challenges across several data sources.

The initial stage in data cleaning is detecting and handling missing data, which can be accomplished using techniques such as imputation (filling in missing values) or eliminating rows/columns with insufficient data. Next, duplicates are discovered and deleted to prevent repetition and ensure data uniqueness. Errors, such as typos or wrong values, are then repaired, usually with domain knowledge or automated processes. Outliers, which can skew statistical analyses, are also recognized and either deleted or changed according to the analysis objectives.

Data cleaning is necessary because clean data improves the accuracy and dependability of subsequent analysis, modeling, and decision-making. It greatly improves machine learning models' performance by ensuring they are trained on high-quality, representative data, resulting in more actionable insights and improved outcomes for businesses and organizations.

4.7.4 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a critical step in the data science life cycle that entails examining and summarizing a dataset's key characteristics, generally using visual approaches. The purpose of EDA is to get a knowledge of the data's structure, patterns, correlations, and anomalies in order to aid future analysis and modeling. It allows data scientists to analyze data quality, discover trends, and find outliers or missing numbers.

Various approaches are employed in EDA, including statistical summaries (mean, median, standard deviation) and visualizations like as histograms, box plots, scatter plots, and correlation matrices. These methods aid in gaining insights into variable distribution, identifying potential feature correlations, and determining the necessity for data preparation. Generating hypotheses is a crucial component of EDA, which motivates data scientists to create possible theories on the connections in the data that may be investigated later. EDA is also crucial for feature selection since it identifies the variables that are most crucial for modeling. In the end, EDA facilitates decision-making by offering a comprehensive comprehension of the data, guaranteeing that further analysis and machine learning models are constructed upon a strong basis.

4.7.5 DATA MODELING AND DATA EVALUATING

Data modeling is an important step in the data science life cycle, in which data scientists use algorithms and statistical approaches to create prediction models based on a prepared dataset. The purpose of data modeling is to identify patterns or relationships in the data that can be used to anticipate or make judgments about future data. This phase typically follows data collection, preparation, and exploratory data analysis (EDA).

Various sorts of models are used during data modeling based on the task at hand. These may include classification models (e.g., decision trees, logistic regression) for categorizing data, regression models (e.g., linear regression) for predicting continuous values, and clustering models (e.g., k-means) for grouping related data points. More complex datasets can also be analyzed using machine learning methods like neural networks and random forests.

The modeling method begins with training the model on a subset of the dataset, followed by evaluating its performance using measures like accuracy, precision, recall, or mean squared error. To boost performance, the model is refined via tuning, cross-validation, and feature engineering. Successful data modeling ensures that the model is resilient, generalizable, and capable of producing trustworthy predictions when deployed, resulting in actionable insights and informed decision-making across sectors.

Data evaluation is a vital phase in the data science life cycle that assesses the performance of models developed during the data modeling phase. The goal of data

evaluation is to ensure that the model fits the training data well while also generalizing effectively to new, previously unknown data. This step helps to verify whether the model's predictions are accurate, dependable, and appropriate for making informed judgments.

During the evaluation phase, many methodologies and indicators are employed to measure model performance. In classification tasks, metrics including as accuracy, precision, recall, F1 score, and confusion matrices are used to assess how successfully the model distinguishes between classes. Metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared are used to evaluate how effectively a regression model predicts continuous data.

Cross-validation is frequently used to evaluate the model's performance across different subsets of the dataset, lowering the risk of overfitting. If the model performs poorly, data scientists may return to earlier stages, such as feature engineering, model selection, or hyperparameter tuning, to enhance it. This iterative approach guarantees that the final model is reliable, accurate, and suitable for use in real-world applications.

4.6.6 DATA DEPLOYMENT AND MAINTENANCE

Data deployment and maintenance are the final steps of the data science life cycle, in which models are put into production and regularly monitored to ensure optimal performance.

Data deployment is integrating a learned machine learning model into current business processes or systems. This could include exposing the model via an application interface (API), integrating it with databases, or deploying it on cloud platforms for real-time predictions. During deployment, the model's capacity to handle live data, scale successfully, and interface with other systems is evaluated to guarantee smooth integration and operation. This step also includes creating automated pipelines to manage incoming data and retraining the model as needed.

Data maintenance is an ongoing procedure that ensures the deployed model continues to function properly over time. Models can decline or become less accurate when the data used to train them changes, a phenomenon known as "model drift." Therefore, maintenance entails frequently assessing the model's performance, updating it with fresh data, and retraining it as needed. It also entails modifying model parameters or upgrading the model as performance deteriorates or the underlying data patterns change.

4.8 CAREERS IN DATA SCIENCE

With the exponential growth of data and the increasing reliance on data-driven decision-making, data science offers high job security, competitive salaries, and opportunities for continuous learning. A profession in data science can be found in a wide range of sectors, such as healthcare, banking, retail, technology, and entertainment. For example, data scientists support personalized treatment and predictive diagnoses in the healthcare industry, and they optimize marketing and supply chain tactics in the retail industry. Ascending to positions like senior data scientist, machine learning engineer, or data science manager, the career ladder usually begins with jobs like data analyst or junior data scientist. Chief data officer and AI research scientist are examples of advanced positions.

Data Analyst : Data analysts play a crucial role in sectors where accurate data interpretation is essential for decision-making, such as technology, marketing, retail, healthcare, and finance. Data analysts may, for instance, evaluate risk or find investment opportunities in the financial industry. To improve strategy, marketers assess consumer behavior and campaign performance. Analysts make contributions to the healthcare industry by researching resource allocation or patient outcomes.

In finance, data analysts evaluate investment trends, track risks, and optimize financial plans. In healthcare, they examine patient outcomes and optimize resource allocation, resulting in improved care delivery and operational procedures. In retail, data analysts are critical for inventory management, consumer segmentation, and sales forecasting. The IT industry relies on them for product analytics, user experience enhancement, and software performance measures. Furthermore, government and non-profit groups employ data analysts to address societal concerns, streamline services, and improve policy outcomes.

Data Scientist : The career scope of a data scientist in the field of data science is expansive and continues to grow as organizations increasingly leverage data to gain insights, make decisions, and innovate. Data scientists play an important role in solving complicated problems by evaluating huge, unstructured databases, developing prediction models, and deploying machine learning algorithms. Their work is spread throughout several industries, including technology, healthcare, finance, retail, entertainment, and manufacturing. For example, they create recommendation engines for e-commerce platforms, optimize financial risk management methods, improve diagnostics and individualized treatments in healthcare, and enable predictive maintenance in manufacturing.

As data becomes an increasingly valuable asset for businesses, the career opportunities for data scientists are likely to expand tremendously. With the rise of automation, IoT, and sophisticated analytics, data scientists will play an important part in determining the future of technology and business, making this career path both financially rewarding and intellectually stimulating.

Data Engineers : Data engineers are key experts in data science who design, create, and maintain the infrastructure and systems that allow firms to gather, store, and process huge amounts of data efficiently. They prioritize ensuring that data pipelines and infrastructures are strong, scalable, and optimized for analysis and decision-making. Data engineers are responsible for developing, constructing, and maintaining the systems and infrastructure required to efficiently manage enormous amounts of data. They are in charge of developing and optimizing data pipelines that gather, analyze, and convert raw data into formats that can be used by data scientists, analysts, and business applications. Data engineers create and manage data architectures including data warehouses, lakes, and databases to ensure scalability, dependability, and performance. They use ETL (Extract, Transform, and Load) processes to clean and prepare data for analysis while maintaining quality and consistency.

Machine Learning Engineers : The career scope of a machine learning engineer in data science is expansive and rapidly growing, as businesses across all sectors increasingly recognize the value of machine learning in solving complex problems and driving innovation. Machine learning engineers are experts in creating, implementing, and

deploying machine learning models, which enable computers to learn from data and make predictions or judgments without explicit programming. They serve as a liaison between data scientists who create algorithms and software engineers who put scalable systems into action. Their work is critical for applying machine learning to real-world challenges, making them indispensable in fields ranging from technology and healthcare to finance and logistics.

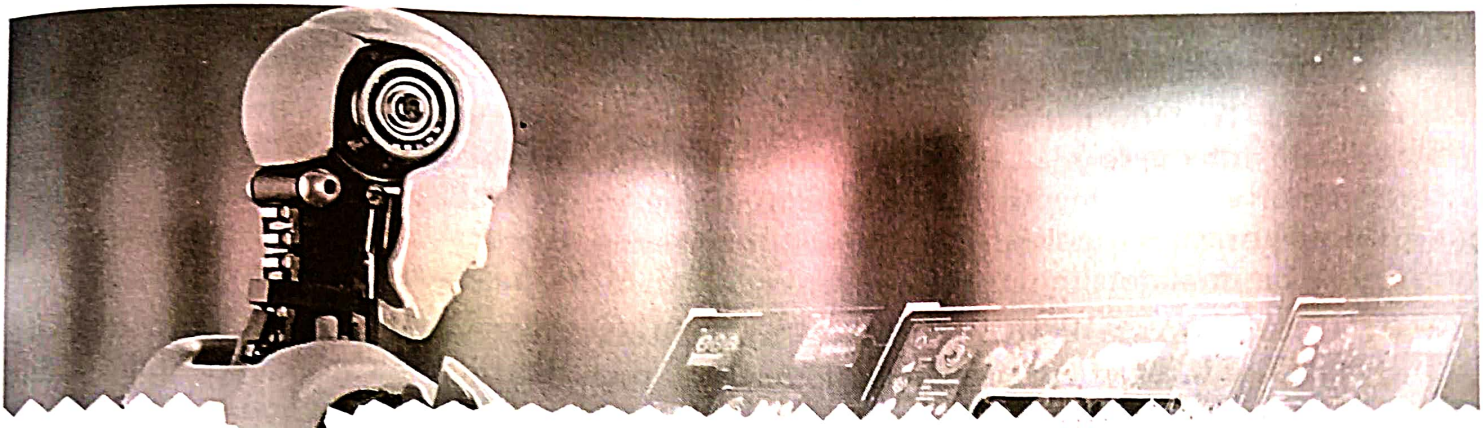
Machine learning engineers play an important role in the creation and deployment of machine learning models, which allow computers to learn from data and make intelligent decisions autonomously. Their key tasks include developing, training, and maintaining machine learning models to achieve specific business goals. They choose and use relevant algorithms and techniques, such as supervised learning, unsupervised learning, and deep learning architectures, to tackle complicated issues in fields such as healthcare, finance, and technology. Machine learning engineers preprocess and clean data to make it appropriate for training models. This includes resolving missing values, standardizing features, and developing new features that increase model performance.

It is the responsibility of machine learning engineers to deploy models into production environments after they have been developed and tested. Using cloud services like AWS, Google Cloud, or Azure, as well as tools like Docker and Kubernetes to make sure models can scale effectively, this entails building up the required infrastructure. They also concentrate on preserving model performance by retraining models as needed, solving problems like model drift, and regularly monitoring and updating models to adjust to shifting data and business circumstances. Working together is essential; machine learning engineers collaborate extensively with data scientists to align their models with business objectives and with software engineers to guarantee smooth integration with current systems.

QUESTIONS

1. Define data science. Explain the evalution of data science.
2. Explain the need of data science.
3. Explain the various types of data used in data science.
4. What are the various fields related to Data Science? Explain.
5. Explain various applications of data science.
6. Explain data science life cycle.
7. Explain various careers in data science.

☆☆☆☆☆



CHAPTER

5 DATA PRE-PROCESSING

CHAPTER AT A GLANCE

5.1	Introduction	5.5	Data Integration
5.2	Why is Data Preprocessing Important?	5.6	Exploratory Data Analysis
5.3	Major Tasks in Data Preprocessing	5.7	Data Reduction
5.4	Data Cleaning	5.8	Data Storage and Management

5.1 INTRODUCTION

Data pre-processing is an essential phase in the life cycle of data science and machine learning. It involves converting unprocessed data into a format suited for modelling and analysis. The aim of data pre-processing is to enhance the quality and use of the data, which eventually improves model performance, data analysis, and decision-making. Raw data is frequently noisy, inconsistent, and incomplete in real-world situations. Machine learning models may yield erroneous or untrustworthy results if they are not properly preprocessed. Therefore, data preparation guarantees that the data is organized, clean, and prepared for additional analysis. Following are some of the main tasks in data pre-processing:

- **Data Cleaning :** The process of identifying and handling mistakes, inconsistencies, and missing values in a dataset is known as data cleaning. For instance, there are numerous causes of missing data, including data corruption, problems with data collecting, and mistakes in data entry. Since missing values have a substantial impact on a model's accuracy, they must be addressed. Missing data can be handled in a number of ways, such as by deleting rows or columns that include missing values, filling in missing values using statistical techniques, or employing algorithms that can deal directly with missing values.
- **Data Integration :** Data from various sources is used in many real-world projects, and it might be difficult to combine them into a coherent dataset. Assuring that

all datasets are consistent in format and structure frequently entails merging data from databases, spreadsheets, or APIs.

- **Data Transformation :** Data transformation is the process of transforming raw data into a format that is more suited for analysis or modelling. This could involve switching data types, aggregating or separating columns, and encoding categorical variables. For instance, time-series data may require transformation to accommodate time zones or deal with date format anomalies.
- **Data Normalization and Standardization:** Data normalization and standardization are used to scale features such that they all fall into a comparable range. Normalization rescales feature values within a predefined range, often between 0 and 1, which is particularly useful for models where the scale of features varies greatly. In contrast, standardization centers data around the mean (0) and scales it according to standard deviation (1).
- **Handling Outliers:** Data points that deviate substantially from other observations are known as outliers. They have the ability to distort model predictions and statistical studies. Finding and dealing with outliers, whether by eliminating or altering them, contributes to the accuracy of the model.

5.2 WHY IS DATA PREPROCESSING IMPORTANT?

Data preprocessing is important because it directly impacts the quality, accuracy, and efficiency of any data analysis or machine learning model. Raw data is often messy, incomplete, inconsistent, and noisy, which can lead to misleading results or poor model performance, if used without proper cleaning and transformation. Data preporcessing is important for the following reasons :

- **Eliminating Errors :** Eliminating errors in data preprocessing is critical to assure the quality and reliability of dataset. Errors can come from a variety of causes, including data entry mistakes, system problems, and inconsistencies between data sources. To rectify these problems, first detect them using approaches such as data validation, visual examination, and consistency checks. Typographical errors, such as misspelled categories or wrong numbers (e.g., "1000" instead of "100"), can be repaired by specified rules or pattern-matching algorithms. Inconsistent units or formats, such as different date formats or measurement units, should be standardized to a single norm. Duplicate records are another major cause of errors that can be deleted to prevent unbalanced analysis. By systematically finding and fixing these flaws, the dataset becomes cleaner and more trustworthy, which improves both the performance of machine learning models and the quality of any insights produced from the data.
- **Improves Data Quality:** Improving data quality during preprocessing is critical to ensure the accuracy, consistency, and dependability of dataset for analysis or modelling. This procedure entails a number of critical responsibilities, including processing missing numbers, rectifying errors, and reducing discrepancies. Missing data can be resolved using approaches such as im

median, or predictive models) or by eliminating impacted entries if they are not critical. To ensure consistency, data errors are found and fixed, such as typographical errors or mismatched units. Outliers, which can skew statistical analysis, are identified and either rectified or deleted. Furthermore, data quality is improved by standardizing and normalizing numerical features to ensure uniformity throughout the dataset, making it acceptable for machine learning techniques. Ensuring data integrity improves insights, minimizes biases, and improves model performance, resulting in more trustworthy decision-making. It ensures that data is free of mistakes, inconsistencies, and missing values.

- **Identifying Hidden Patterns :** Identifying hidden patterns in data during preprocessing entails gaining useful insights that may be used to derive subsequent analysis and model creation. This step is critical for detecting correlations, patterns, or abnormalities in the data that may not be immediately obvious. Exploratory Data Analysis (EDA) is a technique that uses statistical methodologies and visual tools such as scatter plots, histograms, and heatmaps to visually represent data distributions and identify relationships between variables. Clustering methods, such as k-means, can combine comparable data points to uncover underlying structures or segments in the dataset.
- **Enhances Model Performance:** One of the most important steps in making sure machine learning algorithms can learn efficiently and produce correct predictions is to improve model performance through data preprocessing. The model's performance is directly influenced by the quality of the data, and preprocessing aids in dataset optimization for improved outcomes. Preprocessing guarantees that the data is accurate and clean by addressing problems including missing values, outliers, and inconsistencies. By ensuring that numerical features are on a comparable scale, techniques like normalization and standardization keep some features from taking center stage in the model because of their greater magnitudes. In addition, dimensionality reduction eliminates extraneous features and noise, allowing the model to concentrate on the most crucial data. By supplying a clean, well-structured dataset, these preprocessing procedures not only expedite the training process but also aid in enhancing model correctness, generalizability, and robustness.
- **Reduces Complexity:** Reducing complexity in data preprocessing removes redundant information from datasets, simplifying them. To make analysis and modelling easier, the dataset must be simplified while maintaining vital information. Raw data may contain irrelevant information, or redundant variables, which can complicate the modeling process and result in overfitting or increased computational costs. To reduce complexity, feature selection is used to identify and keep only the features that are most important to the task at hand. This can be accomplished using approaches such as correlation analysis or feature importance in machine learning models. Dimensionality reduction approaches, such as Principal Component Analysis (PCA), assist in lowering the number of features while retaining the dataset's variance and underlying structure.

Furthermore, data aggregation can be used to group similar data points, summarizing crucial information while lowering granularity. Input preprocessing reduces complexity and ensures that models are not overwhelmed by irrelevant or excessive input, improving both performance and interpretability.

- **Ensures Compatibility:** A crucial step to ensure that the data is in an appropriate format for analysis and modeling is compatibility in data preprocessing. Data frequently originates from a variety of sources with different formats, units, and structures, which might result in discrepancies that make analysis difficult. Preprocessing entails resolving discrepancies between categorical variables (e.g., mapping different labels for the same category), standardizing formats (e.g., making sure all dates are in the same format), and aligning units of measurement (e.g., converting all distances to kilometres) in order to ensure compatibility. Compatibility guarantees that algorithms can handle the data efficiently, avoiding errors and facilitating more seamless model training and evaluation by ensuring that all data points are consistent and adhere to a single structure.
- **Reduces Noise and Outliers** : Reducing noise and outliers during data preprocessing is critical for ensuring that machine learning models identify relevant patterns and make correct predictions. Noise refers to irrelevant or random variations in data that might skew analysis, whereas outliers are extreme values that differ greatly from the rest of the data and have a disproportionate impact on model performance. Noise can be decreased using techniques such as smoothing, filtering, and data aggregation, which help to eliminate unimportant fluctuations. It makes the dataset more representative of the underlying patterns by removing or limiting the impact of noise and outliers, resulting in more robust and accurate machine learning models. This technique helps to prevent overfitting and guarantees that the model generalizes effectively to new data.

5.3 MAJOR TASKS IN DATA PREPROCESSING

Data preprocessing is a crucial phase in data analysis and machine learning that prepares raw data for further exploration and modeling by addressing its imperfections and ensuring it is clean, consistent, and usable. The major tasks in data preprocessing are essential steps to ensure that the raw data is cleaned, transformed, and structured appropriately for analysis or modeling. These tasks help in improving the quality of the data, reducing complexity, and enhancing model performance. The process involves several key tasks that collectively enhance data quality and usability.

- **Data Cleaning:** The first phase is data cleaning, which focuses on correcting mistakes, eliminating inconsistencies, and resolving missing numbers. Missing data can be handled using methods such as imputation (e.g., mean, median, or regression-based techniques) or, if necessary, by eliminating partial entries. Errors such as duplicates, outliers, and inaccurate data entry are detected and repaired to increase reliability.

- **Data Integration:** The next step is data integration, which involves merging data from numerous sources into a single dataset. This process entails resolving naming conflicts, standardizing data formats, and assuring consistency among databases. It also resolves redundancies, such as deleting duplicate records created during the integration process.
- **Data Transformation:** Data transformation involves turning data into an appropriate format for analysis or modeling. This work entails normalizing and standardizing numerical data, encoding categorical variables to make them machine-readable, and altering skewed data to address non-linear correlations. Transformation ensures homogeneity and makes the dataset appropriate for algorithms that require specific input formats.
- **Data Reduction:** Data reduction is another crucial operation that aims to simplify the dataset while retaining valuable information. Techniques like as feature selection (choosing relevant variables), dimensionality reduction (e.g., PCA), and sampling are used to reduce computing complexity while keeping critical insights.
- **Data Discretisation:** Finally, data discretization and binning divide continuous data into intervals, making it easier to read and analyze, particularly for classification jobs. These processes improve data quality, minimize complexity, and ensure interoperability with analysis tools and machine learning models, laying the groundwork for precise and relevant findings.

5.4 DATA CLEANING

The process of systematically locating and fixing mistakes, inconsistencies, and inaccuracies, missing entries from the data within a dataset to increase data consistency and quality in a dataset is known as data cleaning. Data cleaning is also known as data scrubbing. Data cleaning prepares the dataset for analysis and enhancing the accuracy of patterns. Following are some common data errors often found while cleaning data:

- **Duplicate data entries :** Duplicate data entries occur when two or more records in a dataset are identical or very similar. Duplicates can occur for a variety of reasons, including errors during data gathering, integrating datasets from various sources, or repeated data entries. Duplicates in data can result in skewed results, erroneous analysis, and poor model performance.
- **Inaccurate data :** Inaccurate data refers to data that has errors or discrepancies, which can be caused by human error, system flaws, or data entry issues. Inaccurate data can have a substantial impact on the quality of analysis, insights, and machine learning model performance, resulting in unreliable results and possibly wrong conclusions.
- **Missing entries from data :** Missing entries in data refers to the absence of values in specific fields or columns of a dataset. These missing values can occur for a variety of reasons, including errors in data collection, incomplete surveys, and system malfunctions. Missing data management is an important stage in data cleansing since incorrect treatment might result in biased results or poor model performance.

- **Inconsistent data :** Inconsistent data is refers to data that has contradictions, conflicts, or mismatches across several records, fields, or sources. Inconsistent data can come from a variety of sources, including varying data input standards, variances in data collection methods, and conflicts between databases. If not managed effectively, inconsistent data can result in erroneous analysis, misleading insights, and poor model performance.

5.4.1 DATA CLEANING TECHNIQUES

- **Remove inappropriate values:** Datasets may occasionally have redundant or irrelevant information that may not be needed for analysis. In these situations, it is crucial to find and eliminate any records or fields that don't add to the conclusions. The elimination of unnecessary values is one of the major techniques for data cleaning in data mining. Getting rid of unnecessary data from your system should be in top priority. Removing inappropriate values is a crucial step in data cleaning to ensure the integrity and reliability of the dataset. Identifying and addressing outliers, invalid data types, impossible values, duplicates, inconsistent categorical values, and other errors are essential for producing clean and accurate data that supports high-quality analysis and modeling. Each technique should be applied thoughtfully, considering the nature of the dataset and the specific problem at hand. Below are the common types of inappropriate values

  - **Outliers :** Outliers are data points that vary significantly from the rest of the dataset. They can occur due to measurement errors, human errors, sampling erros etc., and they can distort analysis and modeling.
  - **Invalid Data Types** : Inappropriate values can also arise from improper data types, such as numerical columns containing string data or categorical columns containing numeric data.
  - **Negative or Impossible Values** : Certain values might be impossible or illogical in the context of the dataset. For example, a negative age or salary, or a 0 value in a column where it is not possible (e.g., quantity of items sold or number of hours worked).
  - **Duplicate Records** : Duplicate records are redundant entries that can lead to inflated analysis results and incorrect insights. These duplicates might result from repeated data entry, system errors, or integration issues.
  - **Whitespace or Special Characters** : Extra spaces, special characters, or trailing whitespace can interfere with data consistency, particularly for string-based features.
  - **Handling Incorrect or Out-of-Range Dates** : Incorrect or out-of-range dates are problematic for time-based analysis and modeling.
  - **Handling Inconsistent Units** : In datasets that involve measurements (e.g., distance, weight, or temperature), inconsistent units (e.g., kilometers vs. miles or pounds vs. kilograms) can cause confusion.

- **Removing duplicate values :** Removing duplicates is an essential step in the data cleaning process. Duplicates can arise during data collection, merging datasets, or due to errors in data entry. If left unchecked, they can skew analysis, affect the performance of machine learning models, and lead to inaccurate insights. Duplicates can be identified in several ways depending on the dataset and the context as below:
  - **Exact Duplicates:** Records that are completely identical across all fields.
  - **Partial Duplicates:** Records where only specific fields (e.g., customer name and address) are identical, but other fields may differ (e.g., order number).
  - **Near Duplicates:** Records that are almost identical but may have slight variations such as typos, different case usage, or formatting differences.

  Removing duplicates is a critical step in data cleaning that helps improve the accuracy and efficiency of data analysis. Whether dealing with exact duplicates or partial/near duplicates, identifying and removing them ensures that the dataset used for modeling is reliable, consistent, and free of redundancy. Using the right techniques, like leveraging pandas in Python or advanced fuzzy matching, can significantly streamline the data cleaning process and enhance the quality of insights derived from the data.
- **Handle Missing Data :** Missing data is a common issue in real-world datasets and can arise for various reasons such as data entry errors, equipment failures, or incomplete data collection. It is important to address missing data properly because it can lead to biased analysis, incorrect conclusions, or issues with model training. Handling missing data is a critical step in data cleaning to ensure the dataset is complete and reliable for analysis and predictive modeling. The method chosen depends on the nature of the dataset, the proportion of missing values, and the impact of missing data on the analysis or modeling. Simple techniques like mean imputation or removal of rows can be effective when the missing data is minimal. More advanced methods like predictive imputation, multiple imputation, or interpolation are appropriate when the missing data is more complex or prevalent. Careful handling of missing data ensures that the dataset remains robust and ready for accurate analysis and machine learning modeling.

  Here are the key methods for handling missing data:
  - **Removing Missing Data :** One of the simplest approaches to handling missing data is to remove the rows or columns containing missing values. This method is appropriate when the missing data is minimal, and its removal will not significantly affect the overall dataset. Removing missing data is an important step in data cleaning, as it ensures the dataset is complete and ready for analysis or modeling. However, handling missing data requires a careful approach to balance data integrity and the need to retain meaningful information. Simply removing missing data without proper evaluation can result in biased results or loss of valuable data. Removing missing data is appropriate in the following scenarios:

1. **High Percentage of Missing Values:** If a column or row contains a significant proportion of missing data (e.g., >50%), retaining it may contribute little to analysis and can be safely removed.
2. **Non-Critical Data:** If the missing values are in fields that are not essential for the analysis or decision-making process, those rows or columns can be removed.
3. **Unrecoverable Data:** When there are no practical ways to impute missing values or retrieve them from external sources, removal may be necessary.

- **Imputation of Missing Data :** Imputation involves replacing missing values with estimates based on the available data. This is a common technique when dropping missing values would result in losing valuable data. The benefits for imputation are as under:

1. **Preserves Dataset Size:** By filling in missing values, imputation avoids the loss of rows or columns that might otherwise be removed.
2. **Maintains Representativeness:** Ensures that all patterns and relationships in the data are retained, making the dataset more robust for analysis.
3. **Prevents Bias:** By carefully imputing missing data, biases introduced by systematic data loss can be mitigated.
4. **Improves Model Performance:** Machine learning models often require complete datasets. Imputation provides a way to meet this requirement without discarding data.

- **Handling Inconsistent Data :** Handling inconsistent data is an essential part of data cleaning. It involves standardizing data formats, reconciling contradictions, and ensuring that different data sources align with one another. Inconsistent data can arise from conflicting categories and formats to erroneous or missing values. By applying appropriate techniques such as standardization, data validation, and domain-specific corrections, you can ensure that the dataset is clean, consistent, and ready for reliable analysis and modeling.

Here are some common types of inconsistent data

- **Inconsistent Data Format :**Different data sources may use different formats for the same data type (e.g., date formats, currency symbols, or numerical values with commas). These inconsistencies can cause confusion during analysis and may hinder the ability to merge or analyze data effectively.
- **Inconsistent Categorical Data :** Inconsistent categorical data arises when the same category is represented with different labels, spellings, or formats. For example, "yes" might appear as "YES", "yes", "Yes", or "y".
- **Inconsistent or Missing Units :**Inconsistent or missing units can cause confusion, especially for measurements such as weight, length, or currency. If one dataset uses inches while another uses centimeters, it creates inconsistencies.
- **Contradictory Data Entries :** Contradictory data entries occur when a dataset contains values that conflict with each other. For example, an employee's date

of birth might suggest that they are younger than their employment start date.

- **Standardizing Text Data** :Text data in a dataset may vary due to differences in spacing, punctuation, or irrelevant characters, which can create inconsistencies.

- **Forward Fill / Backward Fill** : Forward Fill and Backward Fill are simple and effective techniques for imputing missing values in data cleaning, particularly useful for sequential or time-series data. These methods involve propagating non-missing values forward or backward to fill gaps, ensuring continuity in the dataset.

  Forward Fill replaces missing values with the most recent non-missing value that appears earlier in the data. Forward Fill is used when the missing values represent a continuation of the previous state or trend. It is suitable for time-series data, such as stock prices or sensor readings, where past values logically extend to fill gaps. It is easy to implement and interpret. It retains the temporal or sequential nature of the data.

  Backward Fill replaces missing values with the next non-missing value that appears later in the data. It is used when the missing values represent a continuation of a future state.

5.4.2 DATA CLEANING BENEFITS

Data cleaning is an essential step in the data preprocessing, ensuring that the dataset is accurate, consistent, and ready for analysis. It helps improve the quality of the data, leading to more reliable insights and better decision-making. Here are some of the key benefits of data cleaning:

- **Improved Data Accuracy** : One of the main benefits of data cleaning is the enhancement of data accuracy. By identifying and correcting errors such as incorrect entries, typos, or inconsistencies, you ensure that the data reflects the true nature of the variables it represents. Clean data enables more reliable analysis and reduces the risk of drawing incorrect conclusions.

- **Enhanced Decision-Making** : When data is clean, it provides clearer insights that lead to better business decisions. Inconsistent or erroneous data can lead to faulty conclusions, which may result in poor strategy or operational inefficiencies. Clean data allows organizations to base their decisions on facts rather than assumptions or errors.

- **Increased Efficiency** : Clean data allows for faster processing and analysis. Incomplete, duplicated, or erroneous data can slow down analysis and cause delays in decision-making. By removing redundancies and correcting errors early in the process, you can reduce the time spent troubleshooting and improve the efficiency of data workflows.

- **Better Model Performance** : In machine learning and predictive modeling, the quality of the data directly impacts the performance of algorithms. If the dataset contains missing, duplicate, or inconsistent values, it can affect the accuracy of

the model's predictions. Data cleaning ensures that the data fed into models is of high quality, which in turn improves the model's performance, reduces overfitting, and helps it generalize better to unseen data.

- **Reduced Risk of Legal or Compliance Issues** : In industries like finance, healthcare, and government, data often needs to meet specific compliance and regulatory standards. Inconsistent or inaccurate data can lead to compliance violations, potentially resulting in legal or financial penalties. Data cleaning helps ensure that data is compliant with necessary regulations by removing or correcting errors that could lead to violations.
- **Increased Trust in Data** : When stakeholders (from analysts to executives) work with clean data, they are more likely to trust the results of analyses and the decisions based on them. Clean data reduces uncertainty and helps build confidence in the processes used to derive insights and take action.
- **Improved Data Integration** : In many organizations, data comes from multiple sources. If these data sources contain errors or inconsistencies, it can be difficult to integrate them for analysis. Data cleaning helps harmonize data from different sources, making it easier to combine and analyze them as a cohesive dataset.
- **Cost Savings** : Inaccurate or poor-quality data can result in costly errors. For example, incorrect customer information may lead to failed marketing campaigns, misplaced inventory, or missed opportunities. Data cleaning helps avoid these issues, saving organizations from unnecessary costs and losses.
- **Better Customer Insights** : Clean data provides more accurate insights into customer behavior, preferences, and trends. Accurate and consistent data is crucial for businesses looking to improve customer experience or target marketing efforts. Data cleaning removes inconsistencies, enabling businesses to generate meaningful and actionable customer insights.
- **Facilitates Future Analysis and Reporting** : By cleaning data regularly, organizations can ensure that it remains high-quality for future use. Clean data is easier to work with for ongoing analysis, reporting, and trend identification. As businesses continue to collect and store more data, maintaining data quality becomes essential for long-term success.
- **Better judgment** : Analytics software can function better with more precise data. This makes it possible for businesses to make better judgments about operations and commercial strategy.
- **Improved sales and marketing** : Customer information is frequently inaccurate, inconsistent, or outdated. The efficacy of marketing campaigns and sales initiatives is increased when data in CRM and sales systems is cleaned up.
- **Increased use of Data** : Data is becoming a vital company asset, but without utilization, it cannot produce commercial value. Data cleansing increases the reliability of data, which persuades employees and company executives to use it in their work.

- **Reduced data costs :** Data cleansing prevents problems and inaccuracies in data from spreading further in analytics applications and systems. Because IT and data management teams don't have to keep correcting the same mistakes in data sets, that ultimately saves time and money.

5.5 DATA INTEGRATION

Data integration in data preprocessing is the process of combining data from multiple sources into a unified and cohesive dataset, ensuring consistency and accuracy. This step is essential when data is collected from diverse sources, such as databases, APIs, spreadsheets, or web services, each of which may have different structures, formats, or standards. Proper integration ensures that the combined dataset is coherent and ready for analysis or modeling.

The integration process begins with **schema alignment**, where differences in naming conventions, data types, or structures across sources are reconciled. For example, a column named "Customer_ID" in one dataset might correspond to "Client_ID" in another; data integration ensures these are correctly matched.

Data deduplication is another critical aspect, which involves identifying and removing duplicate records that might result from merging overlapping datasets, thereby improving data quality and reducing redundancy.

Conflict resolution is often required during integration to handle inconsistencies, such as differing units (e.g., kilograms vs. pounds) or contradictory data points (e.g., varying customer addresses). Standardization is applied to ensure all integrated data adheres to a uniform format, making it easier to analyze.

Data integration also involves handling **missing values and gaps** that may arise during the merging process. Techniques like imputation or filling with default values are used to address such issues. Furthermore, integration ensures that relationships between data elements, such as foreign keys in relational databases, are maintained and correctly mapped.

5.5.1 DATA INTEGRATION CHALLENGES

- **Getting Information From Older Systems :** Integrating data from mainframes or legacy systems is arguably the biggest obstacle to data integration techniques. The date and time of the activities, which are typically present in most contemporary systems, are frequently absent from this data.
- **Getting Information From Novel Systems :** These days, there are many new systems that generate various kinds of data from a wide range of sources, including sensors, cloud computing, and Internet of Things devices. Another problem arises since this data may also be unstructured or real-time. Any business that wants to succeed must figure out how to quickly adjust to these new demands.
- **Consistency and Quality of Data :** One of the main challenges in data integration is ensuring data consistency and quality. Organizations frequently run into conflicting information, inconsistent data formats, and different levels of data quality when merging data from several sources. Resolving semantic differences

where same terms may have distinct meanings across systems is just as difficult as addressing formatting inconsistencies.

In order to preserve the integrity of the integrated dataset, addressing missing or partial data also needs careful thought. To guarantee that the final integrated data is correct, dependable, and appropriate for mining, organizations must put strong data cleansing and standardization procedures in place.

- **Data from External Sources :** Any organization cannot rely just on its own internal data to succeed. Organizations must draw from a variety of outside sources to differentiate themselves from their rivals. The majority of these external data sources, however, might not be as detailed or formatted as internal data, which makes integration extremely challenging.
- **Scalability :** Scalability becomes a crucial issue in data integration as data volume, velocity, and variety keep increasing. Integration procedures need to be able to manage ever-larger datasets, frequently in situations that are real-time or almost real-time. This calls for effective algorithms and architectures in addition to strong hardware.

  When new data sources are added, it might be difficult to modify integration procedures, which can greatly raise the system's complexity and resource needs. Maintaining the effectiveness of ensuing data mining operations requires that integration performance grows linearly with data growth.
- **Inappropriate Integration Program :** There is the regrettable risk of selecting the "wrong" kind of software. It can be challenging for enterprises to select the optimal option when there are so many available. Even if you have the correct software, you can be using it incorrectly.
- **Privacy and Security of Data :** Data security and privacy are major issues when integrating data from several sources. To stop unwanted access to private data, it is essential to implement strong access controls for integrated data. In addition, companies must provide data lineage and auditability during the integration process so that the origin and changes of data can be tracked. It takes careful design and execution of security measures to strike a balance between the necessity of complete data integration and the necessity to safeguard sensitive information.
- **Enhancement of Performance :** One of the constant challenges in data integration for data mining is performance optimization. It gets harder to maintain effective query speed as data volumes increase and integration procedures get more complicated. Businesses commonly use caching techniques for frequently accessed data in order to strike a compromise between the data requirement and query response times.
- **Resouce Management:** Another crucial component is resource management, which makes sure that integration duties don't strain system resources and interfere with other processes. As data landscapes and mining requirements change, integration methods must be continuously monitored and adjusted to maintain optimal performance.

- **Schema Integration :** Because different systems employ distinct data types and structures, schema integration is a major difficulty. Carefully mapping attributes between different data sources is necessary to reconcile these discrepancies and produce a single schema. Semantic heterogeneity complicates this procedure since various phrases may represent the same concept across systems or identical terms may have distinct meanings. Furthermore, maintaining schema integration becomes a continuous process as data sources change over time. In order to accommodate source schema changes without interfering with ongoing data mining procedures, organizations must create adaptable integration solutions.

5.5.2 DATA INTEGRATION TECHNIQUES

5.5.2.1 Manual Data Integration

The process of combining the various data sources without any automation is known as manual data integration. Manual Data Integration refers to the process of combining data from different sources or formats through human intervention, without depending on automated tools or algorithms. It typically involves analyzing and merging data manually, ensuring consistency, resolving conflicts, and aligning the structure of datasets. Manual integration is often used when datasets are small, when automated tools are unavailable, or when the complexity of the data requires a more tailored approach. Organization use manual integration for the jobs that need human interaction because of the sensitivity or complexity of the data. The common use cases used by manual data integration are Fraud detection, compliance, and client segmentation.

Advantages of Manual Data Integration

- **Reduced system costs:** Manual data integration does not require the use of costly systems, equipment, or software. When they start out, entrepreneurs usually don't have a large budget. Without investing a lot of money in a powerful integration application, a company might join the market with a manual management system.
- **Reliability:** In a manual integration process, errors are less likely to occur because an employee is keeping an eye on every stage. Employees can typically identify errors as they happen.
- **Greater Precision and Control** : Manual integration allows for a more accurate and controlled process. Human oversight ensures that complex data issues, such as resolving conflicts or understanding the context of discrepancies, can be handled effectively. This level of control is especially useful in situations where automated tools might struggle to make the right decisions due to the complexities of the data.
- **Handling Unstructured Data** :Manual data integration is especially advantageous when working with data that is unstructured, semi-structured, or inconsistent. When data sources are not standardized, automated methods may fail to understand the context or handle errors appropriately. Human intervention can

resolve issues like mismatched column names, inconsistent formatting, and ambiguous data entries that automated systems might overlook.

- **Flexibility in Decision-Making** : Manual integration provides flexibility when decisions need to be made about how to handle conflicting data. For example, if two datasets have contradictory information (e.g., different customer addresses or names), a human can apply domain knowledge or consult external sources to make an informed decision. Automated systems, on the other hand, would require predefined rules and might not be as adaptable.
- **Improved Accuracy in Conflict Resolution** : In cases where data from different sources may have inconsistencies or conflicts (such as data type mismatches or contradictions), manual integration allows humans to carefully resolve these issues. For instance, when merging data from multiple departments, manual intervention can ensure that data conflicts are addressed with proper judgment, whereas automation might apply a one-size-fits-all approach.
- **Custom Tailored Solutions** : Manual integration is ideal when the integration process requires customization. It enables tailored handling of specific data requirements, such as creating custom rules for merging datasets or combining information in ways that automated systems may not easily replicate. It is particularly useful when integrating small datasets or dealing with rare or ideas data situations that require a significant approach.
- **Validation of Data Quality** : Human involvement in the integration process can improve data quality by identifying errors, inconsistencies, or anomalies that automated systems might overlook. Manual review allows for the identification of data entry errors, duplicated records, and irrelevant or outdated information, all of which can be corrected before proceeding with analysis.
- **Suitable for Small or Simple Datasets** : For smaller datasets or when the data is relatively simple, manual data integration can be faster and more efficient than setting up automated tools. It can save time and effort when the volume of data doesn't justify the use of more complex or automated integration methods.
- **No Dependency on Software or Tools** : Manual integration does not require specialized software or automation tools. This can be advantageous when such tools are unavailable, too expensive, or unsuitable for a particular dataset. As long as the person performing the integration has domain knowledge and expertise, they can handle the integration effectively with basic tools like spreadsheets or text editors.
- **Facilitates Better Communication and Collaboration** : When multiple stakeholders are involved in the integration process, manual methods allow for better communication and collaboration. Team members can discuss data issues, agree on standards, and provide input on how best to merge and format the data. This collaborative approach is not as easily achieved through automated systems.

Disadvantages of Manual Data Integration

While manual data integration offers several advantages, it also has notable drawbacks, particularly when applied to large or complex datasets. Below are the key disadvantages of manual data integration:

- **Time-Consuming** : Manual data integration is inherently slow and labor-intensive, especially when dealing with large datasets or multiple sources. The process requires human intervention at each step, such as identifying inconsistencies, aligning schemas, and resolving conflicts. This makes it far less efficient compared to automated tools that can process data much faster.
- **Prone to Human Error** : Since manual data integration relies heavily on human input, there is a significant risk of introducing errors. Mistakes such as incorrect merging of data, misinterpretation of data types, or failure to identify inconsistencies can lead to inaccurate integrated datasets. These errors can compromise the quality of the final dataset and, in turn, the results of any analysis or modeling based on it.
- **Limited Scalability** :Manual data integration becomes increasingly impractical as the volume of data grows. It is suitable for small to medium-sized datasets, but as datasets scale up, the time and effort required to integrate them manually increases exponentially. In large-scale projects, the effort involved in manually reviewing and combining data can become overwhelming, making it difficult to keep up with growing data needs.
- **Inconsistent Results** : Since manual integration involves human decisions, it is more prone to inconsistencies. Different individuals or teams may approach data integration in slightly different ways, which can lead to variations in how the data is combined or cleaned. This inconsistency can affect the integrity of the dataset and lead to unreliable or non-reproducible results.
- **Lack of Automation and Efficiency** : Manual integration lacks the automation and efficiency that modern data integration tools provide. With automated systems, processes like schema mapping, conflict resolution, and data transformation can be completed quickly and with minimal oversight. In contrast, manual methods require continuous input and validation at each stage, which significantly slows down the process.
- **Resource-Intensive** :Manual data integration requires substantial human resources, particularly when large volumes of data are involved. This can strain an organization's workforce, particularly if the team lacks expertise in data management or the required domain knowledge. The resource demands can lead to increased operational costs, particularly in situations where the integration task is repetitive or ongoing.
- **Difficulty in Handling Complex Data Structures** : Manual data integration can struggle with highly complex or unstructured data, especially when there are a large number of data sources with varying formats. For instance, dealing with

data from different systems or databases, each with its own structure, can require deep expertise and take a long time to reconcile manually.

- **Lack of Standardization** : Without standardized processes or tools, manual data integration can result in inconsistent methods being applied across different datasets. There is also a higher chance that the integrated data might not adhere to the same standards or formats, which can create compatibility issues down the line. Standardization across processes is difficult to achieve without automated systems.
- **Limited Documentation and Transparency** : Manual data integration often lacks thorough documentation, making it harder to track how decisions were made during the integration process. This lack of transparency can be problematic when trying to revisit or audit the process later, as there might be no clear record of why certain data conflicts were resolved in a particular way or how data sources were merged.
- **Difficult to Maintain and Update:** As new data sources are added over time, manually integrating them into an already large dataset can be challenging and time-consuming. Updates or changes to the data may also require redoing parts of the integration process, making it harder to keep the dataset current. Automated tools, on the other hand, can handle such updates efficiently, often requiring minimal manual intervention.

5.5.2.2 ETL (Extract, Transform, Load)

ETL stands for **Extract, Transform, Load**, and is a crucial method in data integration that involves three primary steps (Extract, Transform, Load) to consolidate data from multiple sources into a centralized repository, often a data warehouse. ETL enables the efficient extraction of data, transformation into a suitable format, and loading it into a target system where it can be analyzed and used for business intelligence. Here's a detailed explanation of each step in the ETL process:

1. **Extract :** The **Extract** step involves retrieving data from various source systems. These sources could be structured (such as databases, spreadsheets, or ERP systems), semi-structured (e.g., JSON, XML), or unstructured (e.g., text files, emails, web data). The goal of extraction is to collect data without affecting the source systems, often in raw form. Common extraction methods include:
   - **Batch Processing**: Extracting data in scheduled batches, typically during off-hours to reduce impact on the source system.
   - **Real-Time Extraction**: Continuously pulling data as it changes in source systems, often used for time-sensitive data like stock prices or social media feeds.
   - **Incremental Extraction**: Extracting only the new or changed data since the last extraction, minimizing resource usage and improving efficiency.
2. **Transform :** The Transform step is the core of ETL, as it ensures that the extracted data is in a usable format and meets the target system's requirements.

Transformations can involve multiple operations to clean, standardize, and reformat the data. Common transformation tasks include:

- **Data Cleaning**: Handling missing values, correcting errors, and removing duplicates to improve the quality of the data.
- **Data Formatting**: Converting data into a consistent format (e.g., standardizing date formats, converting currencies or units).
- **Data Aggregation**: Summarizing detailed data into higher-level aggregates (e.g., calculating total sales for each month).
- **Data Filtering**: Removing irrelevant or unnecessary data to focus on key insights (e.g., filtering out records older than a certain period).
- **Data Enrichment**: Adding additional data from external sources (e.g., adding geographic information based on a postal code).

The transformation process can be quite complex, especially when data comes from disparate sources with different formats and structures. Therefore, transformation logic often includes rules and mappings that ensure data consistency and accuracy.

3. **Load :** The Load step involves converting the transformed data into the target system, usually a data warehouse or data lake. The data is now in a format that's ready for analysis, reporting, and decision-making. The load process can be done in two ways:

   - **Full Load**: In this method, all the data is reloaded into the target system, replacing the old data with the new dataset. This is typically done when there is no need for incremental updates, or when the data is small and manageable.
   - **Incremental Load**: Here, only the new or changed data (i.e., data that has been added or modified since the last load) is written to the target system. This approach is more efficient for large datasets and ensures that the target system is updated regularly without reloading everything.

   The loading process needs to ensure that the data is inserted correctly into the target system and is available for querying and analysis.

Advantages of ETL in Data Integration

1. **Centralized Data Repository**: ETL allows organizations to consolidate data from multiple disparate sources into a central data warehouse, providing a unified view of the data for reporting and analysis.
2. **Data Quality and Consistency**: The transformation step helps clean and standardize data and hence improving its quality and consistency, which is essential for accurate analysis.
3. **Improved Decision-Making**: With clean, integrated data in a centralized location, businesses can make more informed decisions based on comprehensive reliable information.

4. **Scalability:** ETL processes can be scaled to handle large volumes of data as the organization grows. Batch and incremental loading methods ensure that the system can manage increasing amounts of data efficiently.
5. **Data Transformation Flexibility:** The transformation phase provides flexibility to handle different data structures, allowing for various data manipulation techniques (e.g., aggregation, filtering, and enrichment) based on business needs.

Challenges of ETL in Data Integration

1. **Complexity in Transformation:** The transformation phase can be complex, especially when dealing with data from various sources that need to be cleaned, standardized, and integrated. It requires well-defined rules and thorough testing to ensure data accuracy.
2. **Performance Issues:** ETL processes, especially in batch mode, can place a heavy load on source systems and require significant processing power, particularly when working with large datasets.
3. **Data Latency:** In traditional ETL systems, there may be delays in data availability due to the time needed for extraction, transformation, and loading. This can be a concern for businesses that need real-time or near-real-time data.
4. **Maintenance and Monitoring:** ETL pipelines require ongoing monitoring and maintenance to ensure that they function correctly over time, especially as new data sources and requirements emerge.
5. **Data Loss Risk:** If not carefully designed, the ETL process may inadvertently lead to data loss or corruption, especially during the transformation or loading stages.

5.5.2.3 Software/Middleware Data Integration

In this data integration technique, applications are connected and data is transferred to databases using middleware, or software. Middleware data integration refers to the use of software tools or services that act as an intermediary layer (middleware) between different applications, databases, and systems to enable communication and data exchange. Middleware facilitates the integration of heterogeneous systems by allowing them to connect, share, and exchange data in real-time or batch processes. It abstracts the complexity of communication between systems and ensures that data flows seamlessly across diverse platforms.

Middleware in data integration typically supports message-oriented, event-driven, and service-oriented architectures. By using middleware, organizations can efficiently integrate data from multiple sources, enabling better data consistency and enabling applications to function together more effectively.

It is quite helpful when combining older systems with more recent ones. Features like automatic processing, assured data transfer between the two systems, and real-time synchronization are all possible with middleware solutions. Using middleware, you can easily translate between different formats, which eliminates the need to manually code the conversion process.

Advantages of Middleware Data Integration

- **Seamless Communication Across Diverse Systems** : Middleware allows systems built on different platforms, databases, and technologies to communicate seamlessly. This helps break down silos, enabling organizations to integrate data across legacy systems, cloud services, and modern applications.
- **Real-Time Data Integration** : Message-oriented and event-driven middleware enable real-time data processing and communication between systems. This is especially important for applications that require up-to-the-minute data for decision-making, such as customer relationship management (CRM) and financial services.
- **Scalability** : Middleware platforms are often designed to handle high volumes of data and transactions, ensuring that systems can scale as the data grows. This is particularly important in modern data architectures, where cloud applications, IoT devices, and big data are involved.
- **Decoupling Systems** : Middleware decouples systems, meaning that they don't need to be directly linked or dependent on each other's internal details. This reduces the complexity of system integration and enables organizations to add or modify systems without impacting other parts of the infrastructure.
- **Data Transformation and Standardization** Middleware can handle data transformation and standardization between different systems. This ensures that data from various sources is formatted correctly, cleaned, and integrated according to business rules.
- **Security and Governance** Middleware often includes built-in security features like encryption, authentication, and access control. This ensures that data transmitted across systems is secure and that organizations can enforce governance and compliance policies.
- **Centralized Monitoring and Management** Middleware provides centralized monitoring and logging capabilities, making it easier to track and manage data flows, troubleshoot issues, and ensure data integration processes are running smoothly.

Disadvantages of Middleware Data Integration

- **Experience is necessary:** The middleware must be deployed and maintained by a qualified IT specialist.
- **Restricted capabilities:** Middleware integration may not be possible for some systems.

5.6 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an important component in the data science pipeline that focuses on evaluating and comprehending the structure, patterns, and relationships within a dataset. It involves the use of statistical and visualization techniques

to summarize the dataset's essential properties, discover anomalies, and uncover trends that inform decision-making and lead additional study. EDA is frequently regarded as a link between data gathering and model development, delivering insights that influence the path of the study.

EDA often starts with summary statistics like mean, median, standard deviation, and correlations, which give a picture of the data's distribution and relationships. Visualization tools such as histograms, scatter plots, box plots, and heatmaps are used to detect trends, outliers, and data distributions. For example, scatter plots can highlight correlations between variables, whereas box plots can uncover outliers.

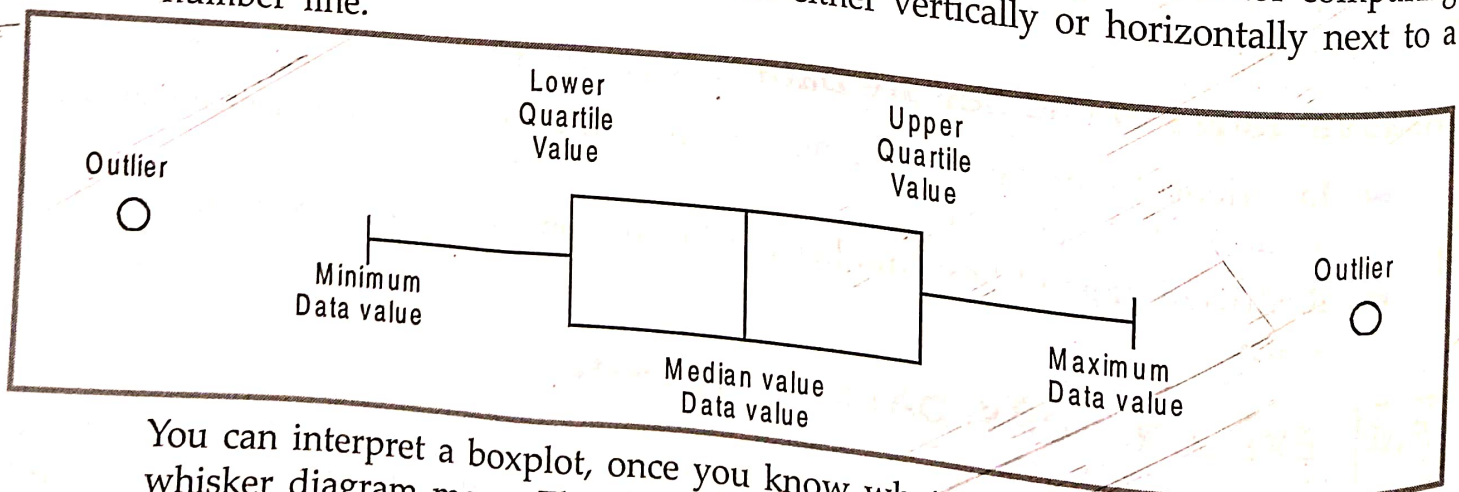
EDA is critical for identifying potential errors in data, such as missing values, discrepancies, or irrelevant features. It also helps with feature selection and hypothesis creation, ensuring that the dataset is thoroughly understood before using machine learning models. Data scientists can use EDA to get key insights, improve data quality, and ensure that future studies or models are correct and meaningful.

5.6.1 TYPES OF EDA TECHNIQUES

Univariate Analysis

The most basic type of data analysis is univariate analysis, in which there is just one variable in the data. "uni" means "one, " there is just one variable in your data. Its primary goal is to describe; it collects data, summarizes it, and looks for patterns in it. Unlike regression, it doesn't deal with causes or relationships. It doesn't address causes or relationships because it only looks at one variable. Finding patterns in the data and describing it are the primary goals of univariate analysis. The main goal of this kind of study is to look at each variable separately. It is usually the first step in any analysis and is the most basic type of EDA in data science. Let's examine some visuals that are utilized to carry out univariate analysis.

- **Quartile :** Quartile is boxplot, sometimes called a box and whisker plot or diagram, is a graph that summarizes a collection of data. The shape of boxplot displays the distribution of the data as well as any outliers. It is a helpful tool for comparing various data sets. These can be shown either vertically or horizontally next to a number line.



You can interpret a boxplot, once you know what the various lines on a box and whisker diagram mean. The **median value** is shown by the line that divides the box in half. This indicates that half of the data is located on the left side of the

median value, while the other half is located on the right. The lower quartile is indicated by the left border of the box, which displays the value that the first 25% of the data falls up to. The upper quartile is indicated by the right border of the box, which displays the value that the 25% of the data is located to the right of the upper quartile value.

- **Histogram:** A histogram is a basic graphical technique for understanding the frequency of variables. It depicts the distribution of data by showing the frequency of occurrence within a range. In a histogram, the attribute under investigation is displayed on the horizontal axis and the frequency of occurrence on the vertical axis. A continuous numeric data type requires a range or binning value to group a set of numbers. For example, in the case of human height in centimeters, all occurrences between 155.00 and 155.99 are classified as 155. There is no optimal bin width or number for all distributions. The distribution becomes more accurate if the bin width is too tiny, but sampling noise becomes visible. Generally, the number of bins should be equal to the square or cube root of the total number of data points.

For example, let we collected data on the heights of 50 students, measured in centimeters (cm). The heights range from 140 cm to 190 cm. The heights are grouped into intervals (bins) of 10 cm, such as 140-149 cm, 150-159 cm, etc.

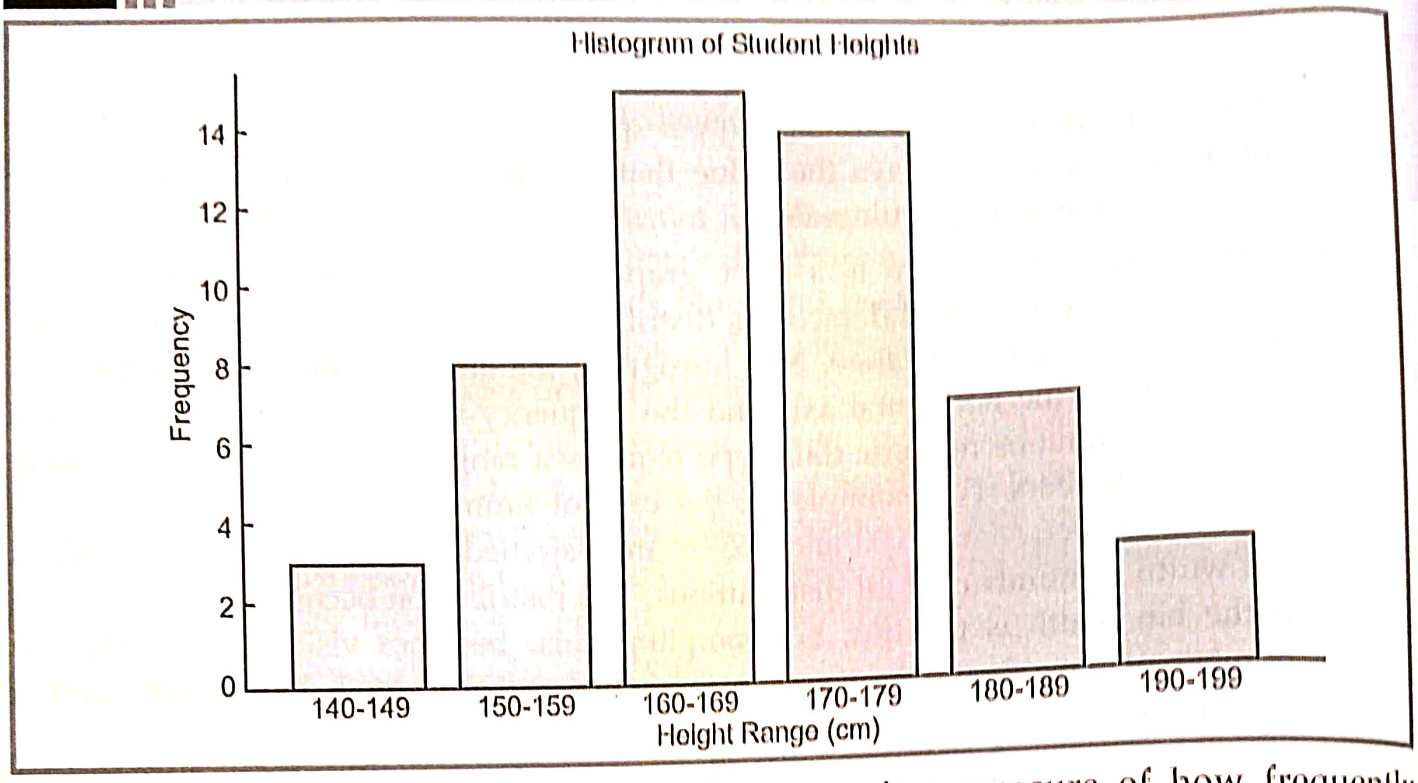
Steps to Create the Histogram:

1. **Data Collection:** Gather heights of the students.
2. **Determine Bins:** Decide intervals like 140-149,150-159, etc.
3. **Count Frequencies:** Count how many students fall into each interval.
4. **Plot:** Create a bar graph where the x-axis represents the height intervals and the y-axis represents the frequency.

Here's how the data might look:

Height Range (cm)	Frequency
140-149	3
150-159	8
160-169	15
170-179	14
180-189	7
190-199	3

Following is a histogram representing the distribution of student heights across different height ranges. Each bar indicates the number of students in a specific height interval.



- **Frequency Distribution Table :** Frequency is a measure of how frequently something happens. For instance, your heartbeat is 72 beats per minute under normal conditions. The frequency of a value is the number of times it occurs. A frequency distribution table presents each data set's frequency in an organized way. It helps us to identify the patterns in the data and allows us to use measures of variance and central tendency to study the data. Making a frequency distribution table out of the gathered data is the first thing a mathematician does with it. Later on, all of the computations, statistical testing, and analysis are completed.

In our daily lives, we encounter a great deal of data in the form of tables, graphs, and numerical numbers. Examples of this data include student grades, city temperatures, points earned in games, and more. Data is the term used to describe the information gathered. After the data is collected, it must be meaningfully represented to make it understandable.

Steps to create Frequency Distribution Table

Step 1: Create a table with two columns: one for frequency and the other for the title of the data you are organizing.

Step 2: Determine whether you want to create a grouped or ungrouped frequency distribution table by looking at the items listed in the data. The grouped frequency distribution table is typically preferable if there are an excessive number of distinct values.

Step 3: Fill in the first column with the data set values.

Step 4: Determine how frequently each item appears in the data that was gathered. Put another way, use counting to determine each item's frequency.

Frequency Distribution Table using Tally Marks

Drawing a frequency distribution table with tally marks is a simple method. For example, if a player tosses the dice, he makes a note of the observations. His observations are as follows: 1, 2, 1, 5, 6, 6, 5, 4, 2, 3, 4, 6, 1, 2, 2, 5. He groups them into categories to determine the precise number of times he received each numeral. The following table shows the concept using tally marks.

Dice Outcome	Tally Marks	Frequency
1	III	3
2	IIII	4
3	I	1
4	II	2
5	III	3
6	III	3

An example of a frequency distribution table is the table above. As you can see, all of the gathered data has been arranged into three columns. A chart that summarizes the values and their frequencies is called a frequency distribution table.

Ungrouped and Grouped Frequency Distribution Table

Ungrouped Frequency Distribution Table: In this table, we record the precise frequency of each piece of data.

For instance, The following are the test grades that 20 students received.

A, A, B, A, C, A, D, E, B, A, B, D, E, A, B, E, A, B, D, E

The table below displays two columns: one for test grades and another for frequency (number of students).

Test Grades	Number of Students
A	7
B	5
C	1
D	3
E	4
Total	20

Grouped Frequency Distribution Table: This type of table is used to organize a lot of data or observations. We create class intervals in order to calculate the frequency of the data that falls within a given class intervals.

For instance, The following are the test scores that 20 students received.

5, 20, 15, 15, 15, 15, 15, 7, 8, 10, 20, 15, 5, 18, 17, 18, 18, 2, 16, 9.

We must create class intervals in order to organize the data in a grouped table. As a result, we will create class intervals with marks ranging from 0 to 5, 6 to 10, and so forth. The table below displays two columns: one for class intervals (test scores) and the other for frequency (number of students).

Test Scores	Number of Students
0 – 5	3
6 – 10	4
11 – 15	6
16 – 20	7
Total	20

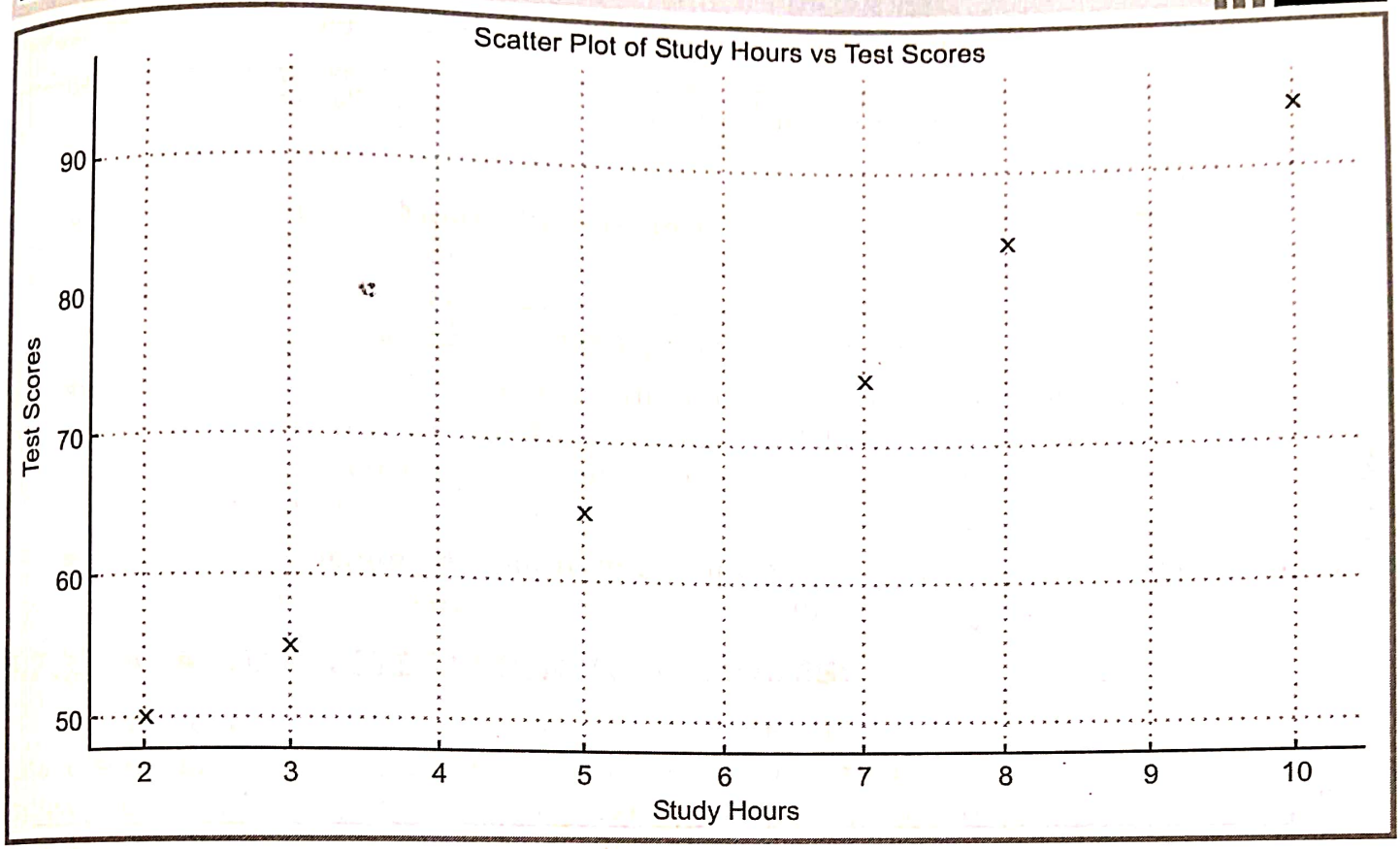
Multivariate Analysis

- **Scatterplot :** Scatterplot is one of the most effective yet straightforward visual displays. The data points in a scatterplot are labelled in Cartesian space with dataset properties that match the coordinates. Typically, the properties are of the continuous data type. Finding a relationship between two attributes under investigation is one of the most important conclusions that can be drawn from a scatterplot. The data points are dispersed if the qualities are not linearly associated, and they align more closely with an imaginary straight line if they are. In addition to showing fundamental correlation, scatterplots can also show outliers in the data and show whether patterns or clusters exist. This is especially helpful for datasets with low dimensions.

  For example, let us take the following data to draw scatter plot representing the relationship between students' study hours and their test scores.

Study Hours	Test Scores
2	50
3	55
5	65
7	75
8	85
10	95

Following is a scatter plot showing the relationship between study hours and test scores. The x-axis represents the number of study hours, and the y-axis represents the test scores. Each point on the graph corresponds to an individual student. Each dot represents a student's data point. This visual helps us observe how test scores tend to increase as study hours increase.



- **Scatter Multiple :** A scatter multiple is an improved version of a basic scatterplot that allows for the simultaneous study of more than two dimensions. For the x-axis coordinate, the primary property is utilized. Additional dimensions or properties are shared with the secondary axis. The colors used in data markers convey the attribute's name. Here, the data points in the highest section of the chart reflect sepal length, the middle section represents sepal breadth, and the bottom section represents petal width. Keep in mind that each property on the y-axis has a duplicate set of data points.

  Here's an example of a scatter plot with multiple groups, representing three classes of students with different study hours and test scores.

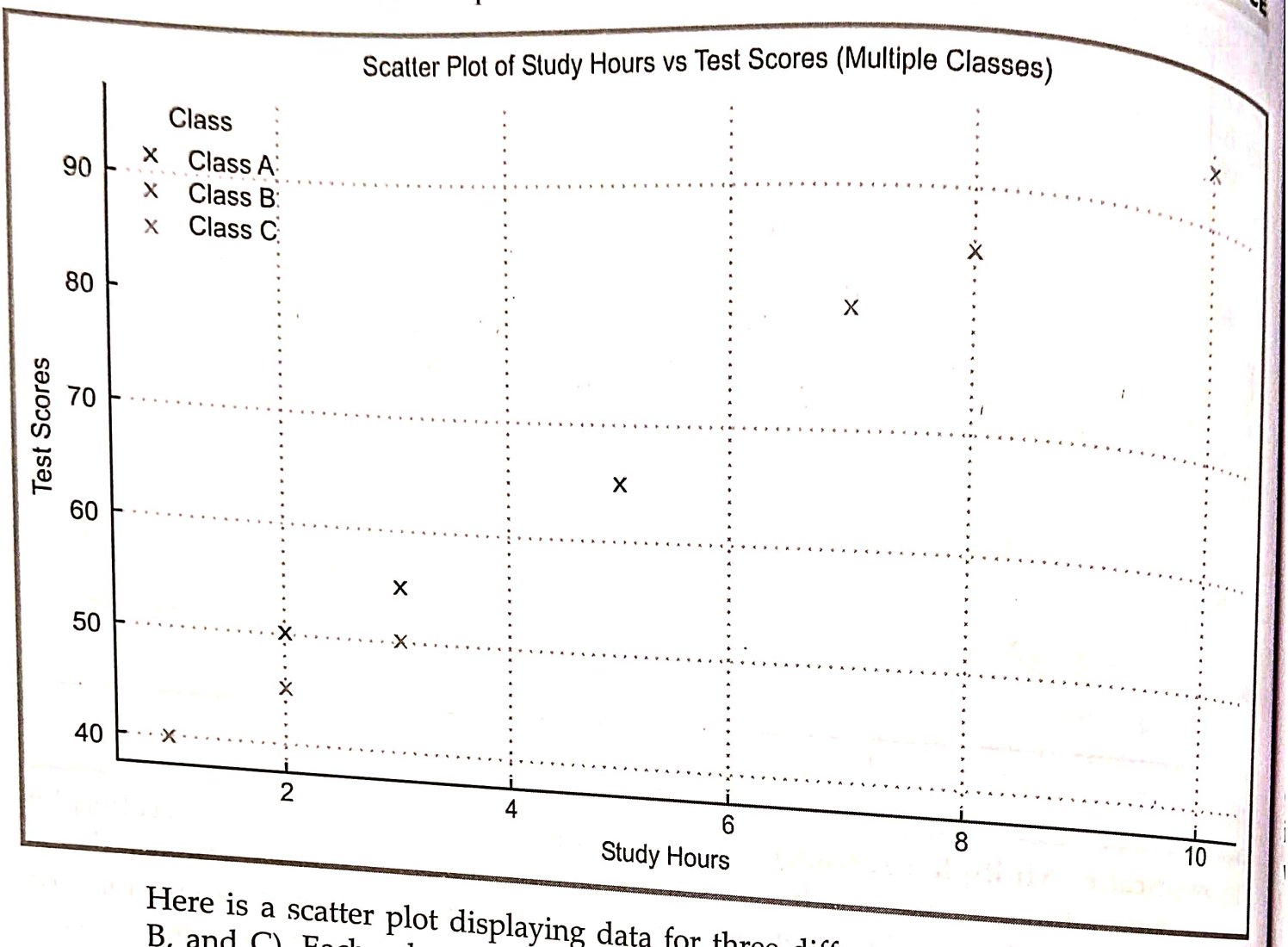
  Description:

  - **Class A:** Students with moderate study habits and scores.
  - **Class B:** Students with lower study habits and scores.
  - **Class C:** Students with higher study habits and scores.

  Example Data:

Class	Study Hours	Test Scores
A	2, 3, 5	50, 55, 65
B	1, 2, 3	40, 45, 50
C	7, 8, 10	80, 85, 95

Let me create a scatter plot to visualize these three groups.



Here is a scatter plot displaying data for three different student groups (Class A, B, and C). Each color represents a class, allowing easy comparison of their study habits and corresponding test scores.

5.7 DATA REDUCTION

In data science, data reduction is the process of reducing the amount of data being examined while keeping the most significant and valuable information. This is accomplished by employing a number of strategies that streamline the data structure, making it simpler to handle, examine, and comprehend. The main objectives of data reduction reducing the computing load, improving model performance, and increasing the effectiveness of data analysis.

5.7.1 DATA DISCRETIZATION

In data science, discretization is the process of breaking down continuous data variables into distinct intervals or bins to facilitate analysis and interpretation. The technique sfrequently employed when working with continuous data that contains a large range of values, such temperature, age, income, or measurement values. By converting data into a more structured format, discretization aids in data simplification, noise reduction, and machine learning model performance enhancement.

5.7.2 PURPOSE OF DATA DISCRETIZATION

- **Reduce Complexity:** Continuous data can be handled and visualized more easily by being divided into a finite set of intervals or bins. This makes analysis simpler and requires less computing power.
- **Enhancement of Model Performance:** When dealing with discrete data instead of continuous variables, several machine learning algorithms operate more effectively. Discretized data can improve the performance of algorithms such as Naive Bayes, k-nearest neighbors, and decision trees.
- **Binning :**Binning is a popular discretization technique that divides continuous variables into bins or intervals. A range of data points is represented by each bin. To produce a categorical variable, for instance, the continuous age data (0–18, 19–24, 25–50, and 51+) can be discretized into age groups.
- **Feature Extraction:** By improving feature extraction, discretized data can make it easier to understand and model the relationships between variables.

5.7.3 DATA DISCRETIZATION TECHNIQUES:

The process of converting continuous data into discrete intervals or bins is known as data discretization. These methods help to improve model performance, simplify data, increase computing efficiency, and make the results easier to interpret. An outline of various popular methods for data discretization is provided below:

- **Equal Width Binning :** In this technique, the data range is divided into intervals of same size using equal width binning, sometimes referred to as range-based or fixed-width binning. A continuous range of data points is present in each interval. For example, if five equal-width bins are created from data ranging from 0 to 100, each bin would span 20 units (0-19, 20-39, 40-59, 60-79, 80-100) . Although this approach is simple and uncomplicated to use, it makes the assumption that data is evenly distributed, which may not be the case in practical situations. It is helpful when there are no notable outliers and the data is evenly distributed.
- **Equal Frequency Binning :** Equal frequency binning, separates data into bins with roughly the same number of data points. This implies that, irrespective of the range, there will be an equal number of observations in each bin. For example, quantile binning would produce 5 bins with 20 data points each in a dataset of 100 data points. This method balances the quantity of data points between bins, producing a more representative analysis, which is advantageous for skewed distributions or data containing outliers. It can assist in preventing the over representation or under representation of particular data items.
- **Clustering using K-means for Discretization :** An unsupervised learning method called K-means clustering divides data into groups according to how similar they are. By putting related data points into bins, it can be applied to discretization. By recognizing organic groups, each cluster serves as a bin, which can streamline complicated data distributions. An iterative procedure is used to reduce the distance between data points and the centroids of each cluster after a number of clusters

(bins) have been chosen. K-means clustering is especially helpful for identifying underlying trends in data and producing more insightful analysis bins.

- **Decision Tree Binning** : Decision tree binning divides continuous data into distinct bins using a decision tree. By representing a split according to a feature's value, each node of the tree efficiently divides data into intervals. This method makes it possible to generate the bins in a hierarchical fashion, capturing significant data splits that show correlations between variables. Decision tree-generated bins typically have meaningful borders, which enhances data interpretability. When working with complicated datasets, it is helpful for producing more complex and context-aware discretization.
- **Entropy Based Discretization** : Information theory is used in entropy-based discretization to quantify and reduce data uncertainty across bins. By grouping data points with high information gain into bins, the goal is to reduce entropy and produce a dataset that is easier to analyze. By choosing the bin limits that preserve the most important information, the technique seeks to improve data quality and lessen noise in the data. This method is a strong tool for classification jobs since it works especially well when the objective is to preserve as much information as possible from the original continuous data.
- **Manual Binning** : Manual binning is a form of discretization where an analyst or domain expert defines the bin ranges based on business logic, expert knowledge, or prior experience. This technique provides high flexibility and customization, but it requires careful consideration and can be time-consuming. It's particularly useful when specific domain knowledge is available, allowing bins to be tailored to specific use cases, such as industry-specific categorization of financial or demographic data.

5.8 DATA STORAGE AND MANAGEMENT

Data storage and management in data science are critical components of any data-driven workflow. They include the procedures for gathering, arranging, preserving, and managing data in a manner that guarantees its security, integrity, and accessibility. Organizations must implement efficient data management techniques in light of the increasing amount and diversity of data. This entails choosing suitable storage options, such as data lakes for mixed data types, relational databases for structured data, and No SQL databases for unstructured data. Scalable and affordable storage alternatives that work well with data science workflows are offered by cloud storage platforms such as AWS S3, Google Cloud Storage, and Azure Blob Storage. To guarantee data quality and get it ready for analysis, data management also entails data integration, preparation, and cleaning.

In order to manage big datasets effectively, distributed file systems, data partitioning, and caching strategies are frequently required. Scalability and performance optimization become more crucial as data quantities increase. Robust analysis, machine learning, and artificial intelligence applications are made possible by first converting raw data into insightful knowledge through efficient data management and storage.

Data storage solutions in data science provide the necessary infrastructure for managing, storing, and accessing data at scale. In order to manage a variety of data kinds, including structured databases, unstructured data, and semi-structured data, enterprises want flexible, scalable, and effective storage solutions as data volumes continue to increase dramatically. The following are a few typical data science data storage options:

5.8.1 RELATIONAL DATABASES

Relational databases are organized storage systems that store data in rows and columns using a tabular style. They are frequently utilized in transactional applications where data integrity is essential and are well-suited for handling structured data with a defined schema. Following are the examples of relational databases:

- **MySQL:** A well-liked open-source relational database renowned for its dependability, speed, and user-friendliness.
- **PostgreSQL** :A robust, open-source relational database with sophisticated capabilities including data integrity and support for complicated data types.
- **SQL Server** :SQL Server is a Microsoft product that is well-known for its robust support for business intelligence and analytics as well as its ability to integrate with other Microsoft technologies.

5.8.2 NoSQL DATABASES

NoSQL databases are more adaptable than conventional relational databases and are made to manage unstructured or semi-structured data. They can extend horizontally across numerous servers and don't need a set schema. Following are some of NoSQL database:

- **MongoDB :** MongoDB is a document-oriented NoSQL database that may be used for a variety of purposes, such as data analytics and content management, because it saves data in documents that resemble JSON.
- **Cassandra:** A highly scalable distributed NoSQL database that is tuned for operational performance and high availability.

5.8.3 DATA LAKES

Large volumes of unstructured, semi-structured, or organized raw data are stored in centralized repositories called data lakes. They provide a variety of data processing techniques and enable flexible data storage. Cloud services like AWS S3 or Google Cloud Storage, along with technologies like Hadoop and Apache Spark, are frequently used to build data lakes. They are perfect for big data analytics and machine learning because they can manage data from various sources without the need for predetermined schemas and are especially helpful for storing massive volumes of data. Followig are some examples of Data lakes.

- **Amazon S3 :** A highly scalable object storage service provided by AWS, suitable for storing and retrieving any amount of data.
- **Azure Data Lake Storage :** A cloud-based solution that supports high-speed analytics on large datasets.

- **Google Cloud Storage:** It provides scalable and cost-effective storage for data science applications, including machine learning and big data analytics.

5.8.4 DISTRIBUTED DATABASES

Distributed file systems enable the storage and processing of large data sets across multiple servers. They ensure high availability, reliability, and performance. Some of the examples of distributed databases are as under:

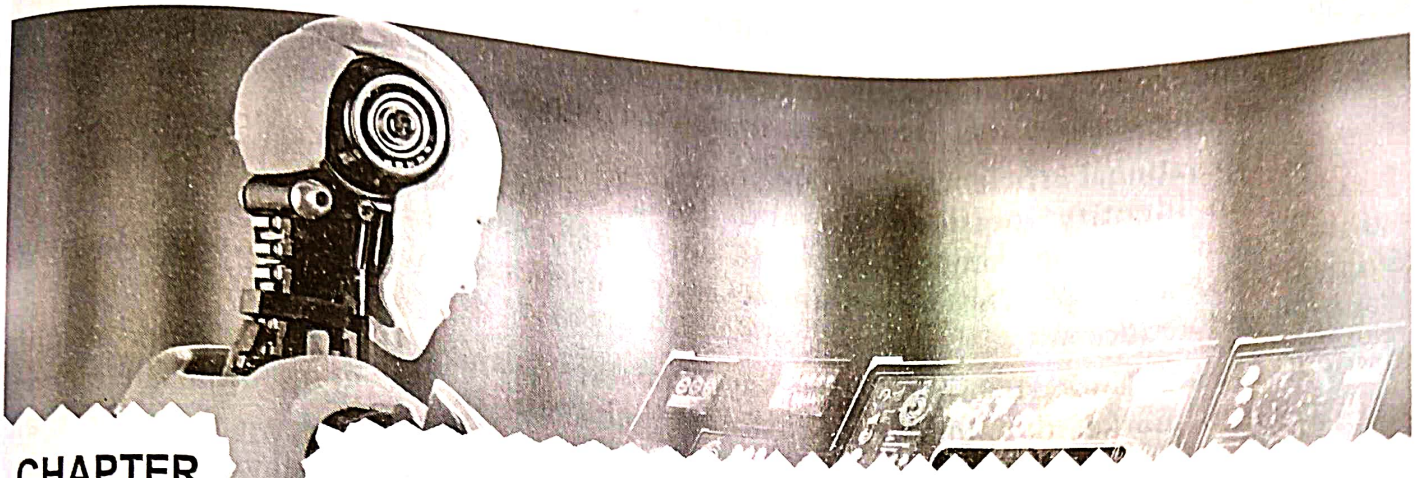
- **Hadoop:** An open-source distributed storage and processing platform that supports large data sets across many nodes in a cluster. It allows for distributed storage of raw data and is often paired with Apache Spark for big data processing.
- **Apache HBase:** A column-oriented NoSQL database built on top of Hadoop and suitable for large-scale real-time data storage.
- **GlusterFS:** An open-source distributed file system that provides scalable and high-performance storage for big data analytics.

5.8.5 CLOUD-BASED STORAGE

Cloud storage solutions provide scalable, flexible, and cost-effective storage options that integrate seamlessly with other data science tools and platforms.They provide strong security, backup, and disaster recovery features and are made to accommodate different data storage requirements. Followig are some examples of Cloud-based storage databases:

- **AWS S3:** An extremely scalable object storage solution that can accommodate a range of data storage requirements, such as analytics and data lakes.
- **Google Cloud Storage:** It offers features like integrated machine learning capabilities along with comparable capability to AWS S3.
- **Azure Blob Storage:** A storage option from Microsoft Azure that can be combined with analytics and machine learning services and accommodates massive volume of data.





CHAPTER

6 SORTING AND FILTERING DATA

CHAPTER AT A GLANCE

6.1 Introduction to Data Analysis	6.3 Data Sorting
6.2 Creating Database	6.4 Filtering Data

6.1 INTRODUCTION TO DATA ANALYSIS

Data analysis is the process of systematically applying statistical and logical techniques to describe, illustrate, condense, recap, and evaluate data. This process involves inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

6.1.1 NEED FOR DATA ANALYSIS

Data analysis is indispensable in today's data-driven world, influencing various sectors and driving numerous aspects of personal and professional decision-making. Below is an in-depth look into the key reasons why data analysis is crucial.

1. **Informed Decision Making:** Data analysis provides a factual basis for decision-making by extracting meaningful insights from raw data. By analyzing historical and real-time data, organizations can identify patterns and trends that inform strategic decisions. For instance, a company might use sales data to determine which products to discontinue or promote based on their performance over time.
2. **Problem Identification and Resolution:** Data analysis helps in pinpointing issues that may not be immediately apparent. For example, by analyzing customer feedback data, a company can identify recurring complaints about a specific

product feature and address the issue. This proactive approach to problem-solving ensures that issues are resolved before they escalate.

3. **Operational Efficiency:** By analyzing workflow and operational data, organizations can identify inefficiencies and areas for improvement. For example, a manufacturing company can use data analysis to identify bottlenecks in its production process and optimize resource allocation to enhance productivity and reduce waste.
4. **Market Understanding:** Data analysis enables businesses to gain deeper insights into market dynamics and customer behavior. By analyzing market trends, competitor strategies, and consumer preferences, companies can tailor their offerings to meet market demand effectively. This helps in staying competitive and responsive to market changes.
5. **Risk Management:** Data analysis is crucial for identifying and mitigating risks. By analyzing historical data, organizations can predict potential risks and devise strategies to manage them. For example, in finance, data analysis can help in assessing credit risk, detecting fraudulent activities, and making informed investment decisions.
6. **Customer Insights and Personalization:** Understanding customer behavior and preferences is essential for delivering personalized experiences. Data analysis enables businesses to segment their customer base, identify preferences, and tailor marketing strategies accordingly. For instance, e-commerce platforms use data analysis to recommend products based on past purchase behavior, thereby enhancing customer satisfaction and loyalty.
7. **Innovation and Development:** Data analysis drives innovation by uncovering new opportunities and areas for development. By analyzing data from various sources, companies can identify gaps in the market and develop new products or services to fill those gaps. This approach fosters continuous improvement and keeps the business ahead of the curve.
8. **Cost Reduction:** Efficient data analysis can lead to significant cost savings. By identifying areas of overspending and inefficiencies, organizations can optimize their operations and reduce costs. For example, in supply chain management, data analysis can help in optimizing inventory levels, reducing holding costs, and improving logistics.
9. **Performance Measurement:** Data analysis is vital for measuring and evaluating performance against set goals and objectives. By analyzing performance data, organizations can assess the effectiveness of their strategies and make necessary adjustments. This continuous evaluation process ensures that the organization remains on track to achieve its goals.
10. **Compliance and Reporting:** In many industries, compliance with regulations is mandatory. Data analysis helps organizations ensure that they comply with relevant laws and standards. By maintaining accurate records and generating compliance reports, companies can avoid legal issues and penalties.

6.1.2 USES OF DATA ANALYSIS

Data analysis is a versatile tool that plays a crucial role across various sectors. Below some common domains are discussed

1. **Business Intelligence and Strategy:** Data analysis is pivotal in business intelligence, enabling companies to transform raw data into actionable insights. These insights guide strategic decisions, helping businesses stay competitive.
   - **Performance Metrics:** Track key performance indicators (KPIs) to evaluate company performance.
   - **Market Analysis:** Understand market trends, customer preferences, and competitive landscape.
   - **Forecasting:** Predict future trends and behaviors, aiding in proactive decision-making.
   - **Customer Segmentation:** Identify distinct customer groups for targeted marketing and personalized services.
2. **Healthcare:** In healthcare, data analysis is used to enhance patient care, optimize operations, and support medical research.
   - **Patient Care:** Analyze patient data to improve diagnosis, treatment plans, and patient outcomes.
   - **Operational Efficiency:** Streamline hospital operations, reduce wait times, and manage resources effectively.
   - **Medical Research:** Support clinical trials and research by analyzing large datasets to find patterns and correlations.
   - **Public Health:** Monitor disease outbreaks and develop strategies for disease prevention and control.
3. **Finance:** Data analysis in finance is crucial for risk management, fraud detection, and investment strategies.
   - **Risk Management:** Identify and mitigate financial risks through predictive analytics.
   - **Fraud Detection:** Detect unusual patterns and anomalies to prevent fraudulent activities.
   - **Investment Analysis:** Evaluate the performance of stocks, bonds, and other investments to guide investment decisions.
   - **Budgeting and Forecasting:** Analyze past financial data to forecast future budgets and financial needs.
4. **Marketing:** Marketing relies heavily on data analysis to understand consumer behavior and optimize campaigns.
   - **Customer Insights:** Analyze customer behavior, preferences, and feedback to tailor marketing strategies.

- **Campaign Optimization:** Measure the effectiveness of marketing campaigns and adjust strategies for better results.
- **Personalization:** Deliver personalized content and offers to different customer segments.
- **Market Trends:** Monitor and predict market trends to stay ahead of competitors.

5. **Education:** Educational institutions use data analysis to improve student outcomes and enhance administrative processes.
   - **Student Performance:** Track and analyze student performance to identify areas needing improvement.
   - **Curriculum Development:** Use data insights to develop and refine educational programs and curricula.
   - **Resource Allocation:** Optimize the allocation of resources such as faculty, classrooms, and materials.
   - **Enrollment Management:** Analyze enrollment trends to forecast future student numbers and plan accordingly.
6. **Government:** Governments use data analysis to inform policy-making, improve public services, and enhance national security.
   - **Policy Development:** Analyze social, economic, and environmental data to develop effective policies.
   - **Public Services:** Improve the delivery of public services by analyzing usage patterns and feedback.
   - **Crime Prevention:** Use data to identify crime patterns and develop strategies for crime prevention.
   - **Resource Management:** Optimize the allocation of public resources and budget planning.
7. **Supply Chain Management:** Data analysis is essential for optimizing supply chain operations, reducing costs, and improving efficiency.
   - **Inventory Management:** Track and forecast inventory levels to prevent stockouts and overstocking.
   - **Logistics Optimization:** Analyze transportation data to find the most efficient routes and methods.
   - **Demand Forecasting:** Predict customer demand to ensure timely production and delivery.
   - **Supplier Performance:** Evaluate supplier performance and maintain quality control.
8. **Sports and Entertainment:** Data analysis is used in sports and entertainment to enhance performance, fan engagement, and business operations.
   - **Performance Analysis:** Analyze athlete performance data to improve training and game strategies.

- **Fan Engagement:** Use data to understand fan preferences and enhance their experience.
- **Event Management:** Optimize event planning and management through data-driven insights.
- **Content Personalization:** Deliver personalized content and recommendations to viewers and fans.

6.1.3 ADVANTAGES OF DATA ANALYSIS

Data analysis offers a multitude of advantages across various sectors, driving efficiency, innovation, and strategic decision-making. Here's a detailed exploration of its benefits:

1. **Improved Decision Making**
   - **Insight-Driven Decisions**: Data analysis transforms raw data into actionable insights, enabling decision-makers to base their strategies on factual information rather than intuition or assumptions. This leads to more accurate and reliable decisions.
   - **Predictive Analysis**: By utilizing historical data and statistical models, businesses can forecast future trends and behaviors. This predictive capability helps in proactive decision-making, allowing companies to stay ahead of market changes and consumer demands.
   - **Real-Time Data**: Modern data analysis tools provide real-time data processing and visualization. This immediacy helps organizations to react quickly to emerging trends or potential issues, maintaining a competitive edge.
2. **Increased Efficiency**
   - **Process Optimization**: Data analysis identifies inefficiencies within business processes. By pinpointing bottlenecks and redundant steps, organizations can streamline operations, reduce waste, and enhance overall productivity.
   - **Resource Allocation**: It helps in optimal allocation of resources by identifying areas that require more attention and investment, and those that can operate with fewer resources. This ensures that resources are utilized effectively, minimizing costs and maximizing output.
   - **Performance Monitoring**: Continuous data analysis allows for ongoing monitoring of performance metrics. This helps in ensuring that operations are running smoothly and goals are being met. Deviations from expected performance can be quickly identified and addressed.
3. **Enhanced Customer Insights**
   - **Understanding Customer Behavior**: Data analysis provides a deep dive into customer behavior, preferences, and purchasing patterns. This understanding enables businesses to tailor their products, services, and marketing efforts to better meet customer needs.
   - **Personalization**: With detailed customer insights, companies can create personalized marketing campaigns, product recommendations, and customer

experiences. Personalization enhances customer satisfaction and loyalty, leading to increased sales and customer retention.

- **Customer Feedback Analysis**: Analyzing customer feedback from various channels (such as surveys, social media, and reviews) helps businesses understand customer sentiments and areas for improvement. This continuous feedback loop is crucial for maintaining high customer satisfaction levels.

4. **Competitive Advantage**
   - **Market Trends Identification**: Data analysis helps in identifying and understanding market trends, providing insights into emerging opportunities and threats. Companies can leverage these insights to innovate and adapt quickly, staying ahead of competitors.
   - **Benchmarking**: Businesses can compare their performance against industry standards and competitors using data analysis. This benchmarking helps in identifying strengths and weaknesses, and devising strategies to outperform competitors.
   - **Strategic Planning**: Data-driven strategic planning ensures that business goals are aligned with market realities and opportunities. This strategic alignment is key to long-term success and sustainability.
5. **Risk Mitigation**
   - **Early Risk Detection**: Data analysis helps in early identification of potential risks by monitoring various risk indicators and anomalies. Early detection enables businesses to take preventive measures and avoid potential losses.
   - **Fraud Detection**: In sectors like finance and e-commerce, data analysis is crucial for detecting fraudulent activities. By analyzing transaction patterns and behaviors, companies can identify and mitigate fraud risks.
   - **Compliance**: Ensuring compliance with regulatory requirements is critical for many industries. Data analysis helps in monitoring compliance and identifying any deviations, thus avoiding legal penalties and maintaining reputation.
6. **Cost Reduction**
   - **Operational Efficiency**: By optimizing processes and eliminating inefficiencies, data analysis helps in reducing operational costs. This can significantly impact the bottom line, especially in industries with thin profit margins.
   - **Supply Chain Optimization**: Data analysis can optimize supply chain operations by forecasting demand, managing inventory, and improving logistics. This reduces wastage and ensures that resources are available when and where they are needed, reducing holding and transportation costs.
   - **Marketing Efficiency**: Analyzing marketing data helps in identifying the most effective channels and strategies, allowing for better allocation of marketing budgets. This ensures higher returns on marketing investments and reduces unnecessary expenditure.

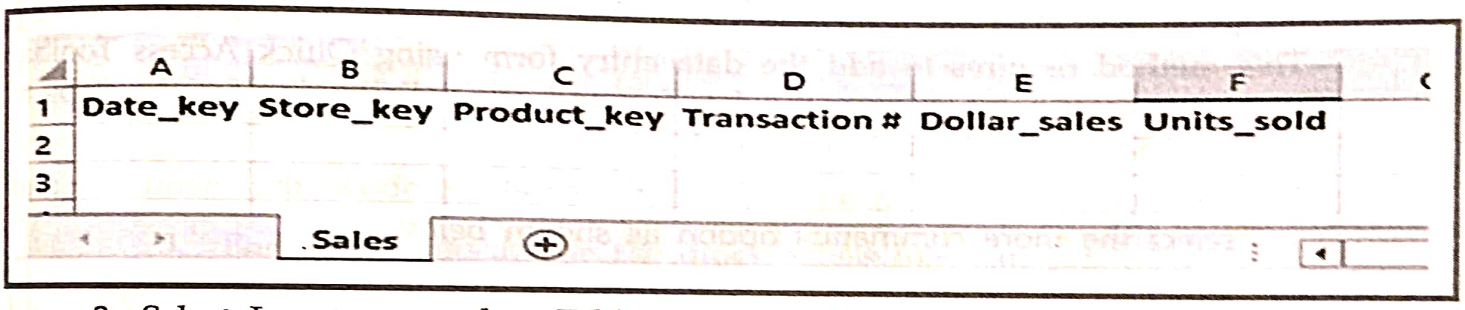
6.2 CREATING DATABASE

Excel workbook contains the excel worksheets and each worksheet contains the rows and columns. Each column can be used as field and each row contains a record. By using row and columns you can build a database in which one can store the data to perform different operations like sorting, filtering, calculation etc. Excel refers to such database as *tables*.

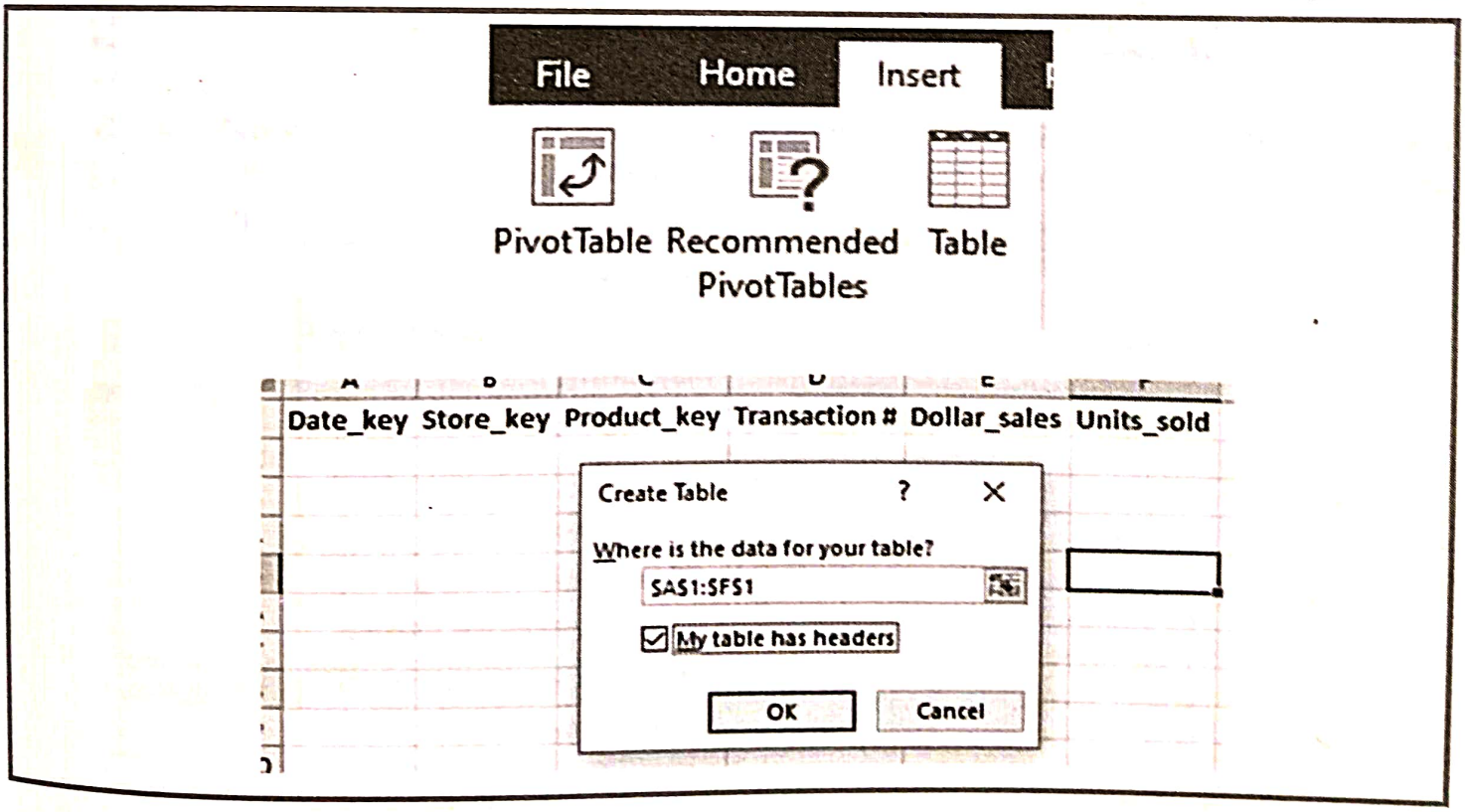
Database can be created either as flat file or relational databases. In flat files all the data is stored in a single table and in relational database, data is stored in different tables whereas each table has one column with unique values and then tables are linked with each other using these unique columns.

Followings are the steps to create the database:

1. Create new workbook in excel using new menu or using short cut Ctrl + N and rename the worksheet with the table name.
2. Type the headings for table in the database as shown below



3. Select Insert menu then Table option to display the create Table dialogue box.



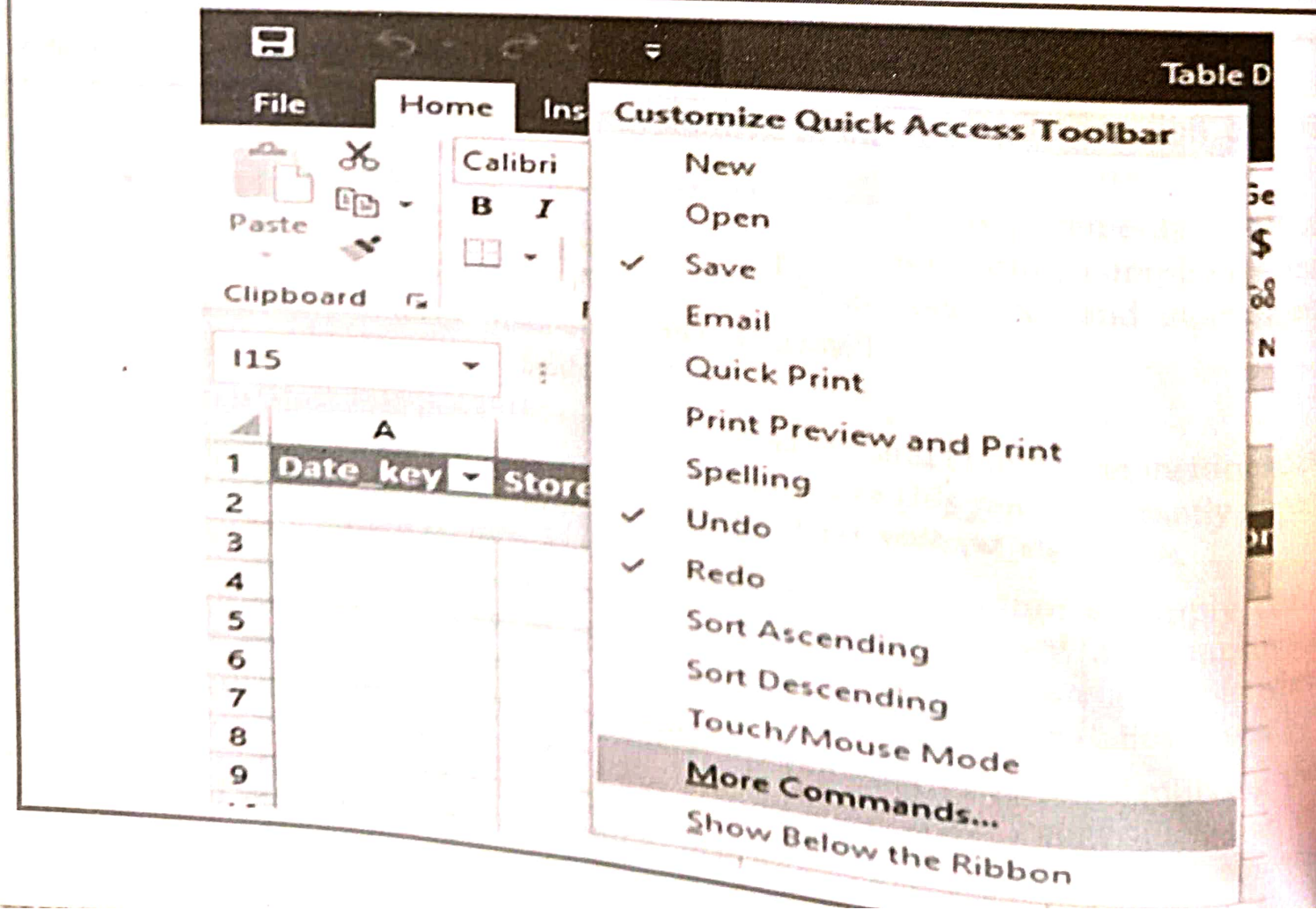
4. Select correct range of your table where data will be entered. Make sure 'My table has headers' option is selected.
5. Click the OK button to close the create Table dialog box. Table is created as shown below:

	A	B	C	D	E	F	G
1	Date_key	Store_key	Product_key	Transaction #	Dollar_sales	Units_sold	
2							
3							

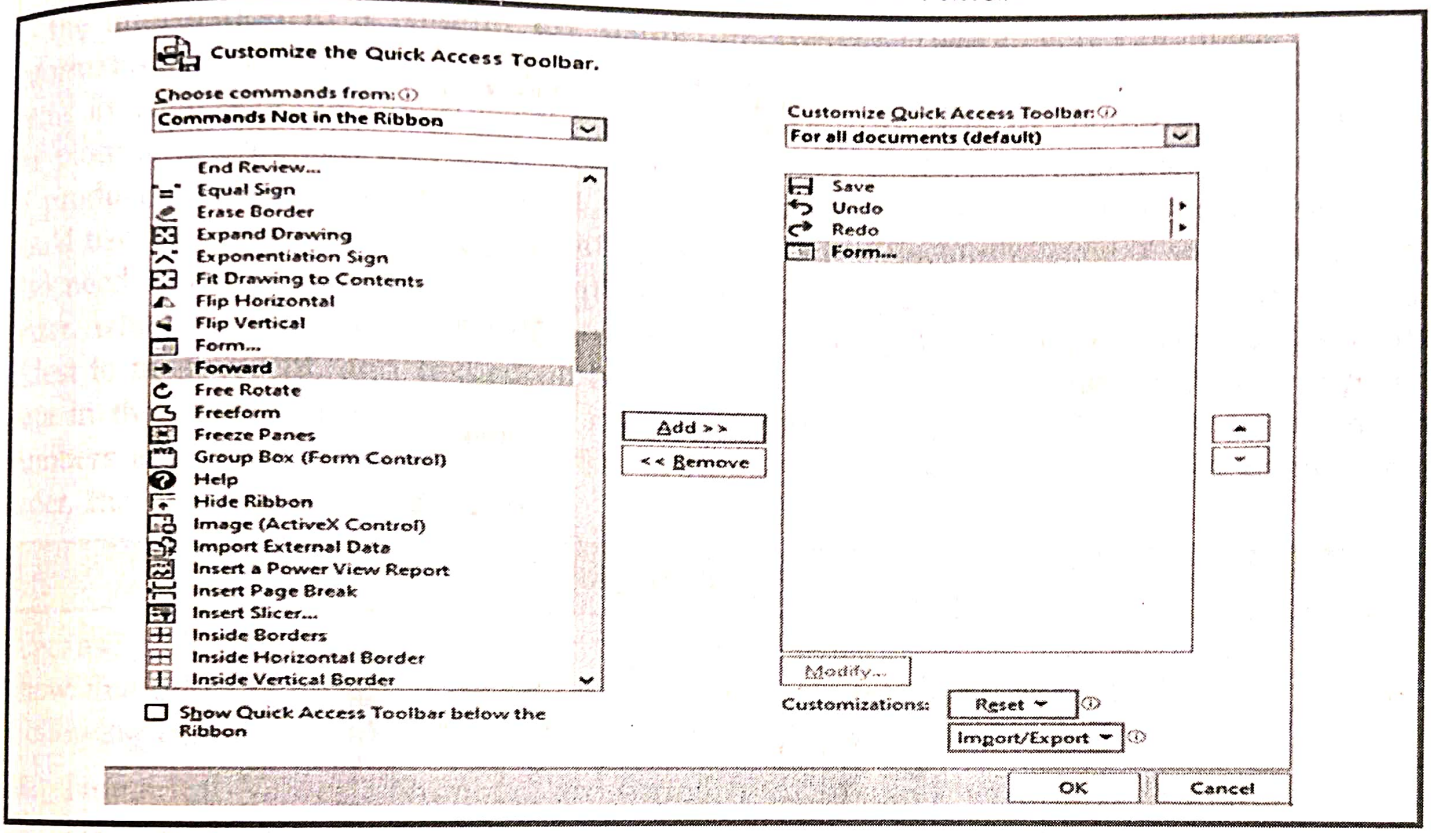
6.2.1 ENTERING DATA IN TABLE

There are two ways to enter the data in a database:

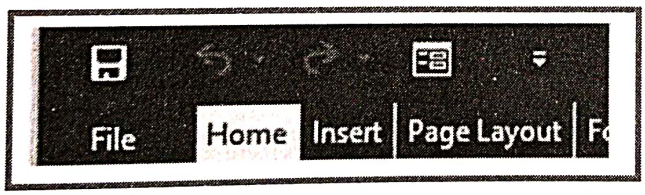
1. **Enter Data Directly:** You can enter data directly by typing the data in the target cell.
2. **Entering Data Using a Data-Entry Form:** When large data is to be entered in the excel then Data entry form is easiest way as compare to enter data directly. This method requires to add the data entry form using Quick Access Toolbar first. This form helps to enter the data fast and easily. Following are the steps to add the data entry form in the excel:
   1. Click on the arrow at top right of the excel i.e. above the menu bar and select the more commands option as shown below:



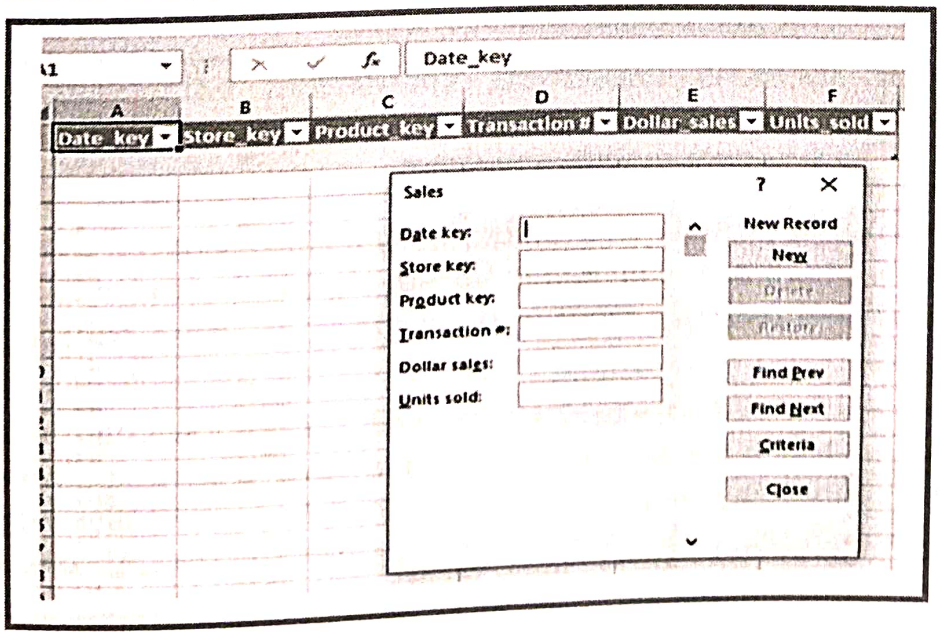
2. Then excel options dialog box appears. Choose 'command not in ribbon' from the 'Choose command from' drop down box. Then select form and click on add button and then click on ok button.



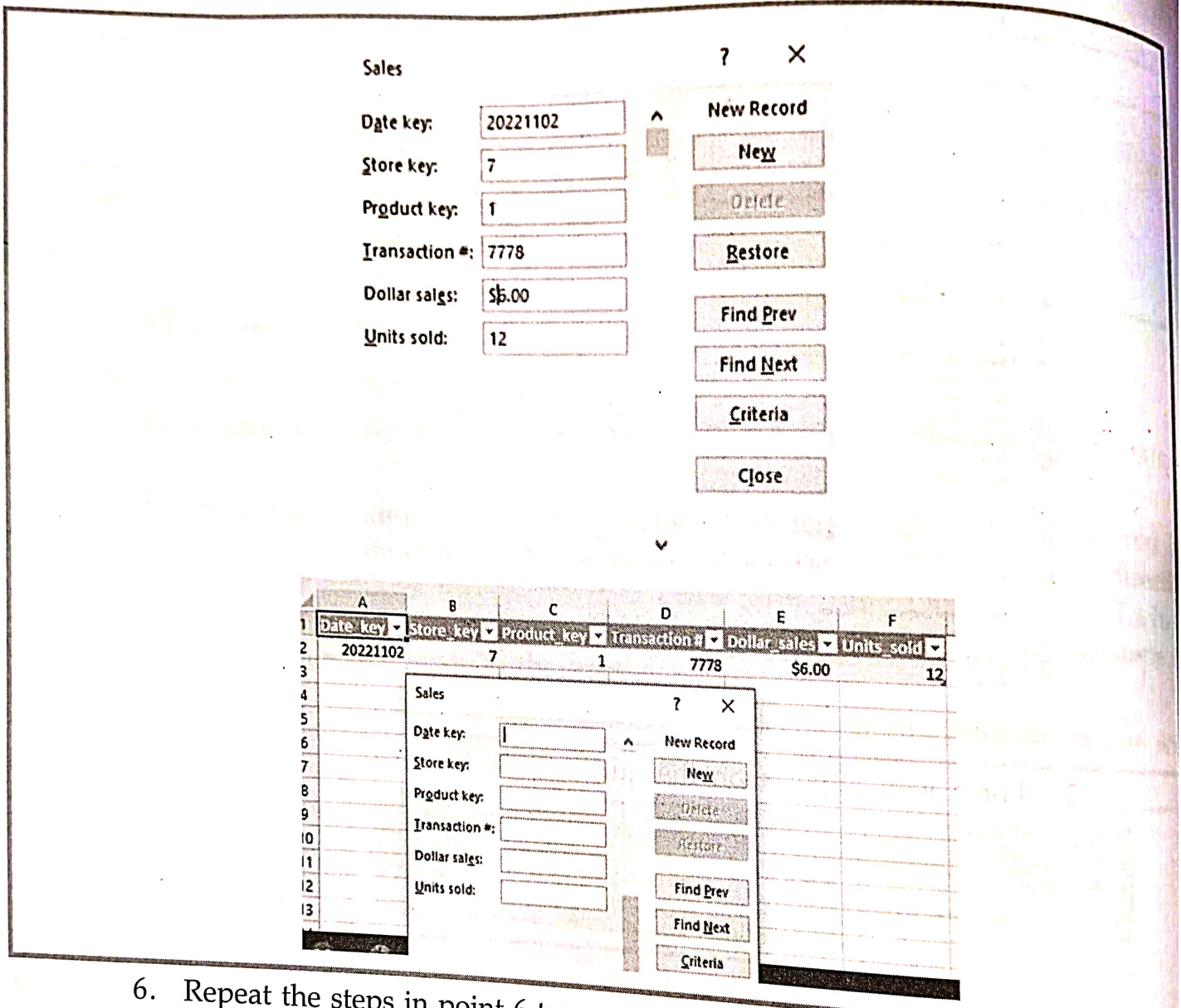
3. Form icon appears on the quick access tool bar.



4. Now add the data in excel table select the first heading of the table and then click on form icon. Form automatically detect the columns of the table and appears as shown below



5. Enter the values in the form and click on new button. It will insert the values in the form and clear the form content for the next value.



6. Repeat the steps in point 6 to enter all records of the table. Click on the clos button when data entry is complete.Following figure shows the sampl records for the SALES database.

Date_key	Store_key	Product_key	Transaction #	Dollar_sales	Units_sold
20221102	7	1	7778	$6.00	12
20221102	7	2	7778	$3.00	1
20221102	7	4	7778	$3.49	2
20221102	7	5	7778	$2.49	35
20221102	7	6	7778	$2.69	2
20221102	7	7	7778	$11.99	1.8
20220906	2	3	5809	$6.00	1
20220906	2	18	5809	$2.48	1
20220906	2	19	5809	$9.81	3
20220906	2	20	5809	$6.54	2
20220906	2	16	5809	$1.70	1.325
20220906	2	21	5809	$4.97	1
20220906	2	22	5809	$3.97	1
20220906	2	3	5809	$5.94	2
20220906	2	23	5809	$2.25	3
20220906	2	24	5809	$1.50	2
20220906	2	2	5809	$1.23	1
20221004	2	25	5809	$53.90	5
20221004	2	15	5809	$31.47	1
	2	50	5625	$5.31	3
		71	5625	$3.97	1
				$2.97	1

6.3 DATA SORTING

Sorting lets you order the items in your table alphabetically or numerically according to the information in a column. By using the correct criteria, you can make sure the information you're interested in appears at the top of the column, making it easier to find items in your table. Before you can sort your data, you need to choose a sorting key—the piece of information that Excel uses to order your records. For example, to sort a table of products so the cheapest (or most expensive) products appear at the top of the table, you'd use the Price column as the sorting key. In addition to choosing a sorting key, you also need to decide whether you want to use ascending or descending order. Ascending order, which is most common, organizes numbers from smallest to largest, dates from oldest to most recent, and text in alphabetical order. (If you have more than one type of data in the same column—which is rarely a good idea—text appears first, followed by numbers and dates, then true or false values, and finally error values.) In descending order, the order is reversed.

Excel's Sorting Rules

Because cells can contain different types of information, you may be curious about how this information is sorted. For an ascending sort, the information appears in the following order :

1. **Numerical Values :** Numbers are sorted from smallest negative to largest positive. Dates and times are treated as values. In all cases, the sorting is done by using the actual values (not their formatted appearance).
2. **Text :** In alphabetical order, as follows : 0 1 2 3 4 5 6 7 8 9 (space) ! ~ # $ % & ' () · +, –./ : ; < = > ? @ [/] ^ - ' {_} – A B C D E F G H I J K L M N O P Q R S T U V W Z Y Z By default, sorting is not case sensitive. You can change this, however, in the Sort Options dialog box (described elsewhere in this chapter).
3. **Logical values :** False comes before True.
4. **Error values :** Error values (such as #VALUE! and #NA) appear in their original order and are not sorted by error type.
5. **Blank cells :** Blanks cells always appear last.

Sorting in descending order reverses this sequence–except that blank cells are still sorted last.

Sorting a table allows you to arrange the data in a specific order based on one or more columns. This can be particularly useful when you want to find the highest or lowest values, alphabetize a list, or group similar items together. Excel provides several methods for sorting tables:

1. **Sort Ascending/Descending :** The simplest way to sort a table is by using the Sort Ascending or Sort Descending buttons on the Excel toolbar. Select the column you want to sort by click on the appropriate button, and Excel will rearrange the data accordingly.

2. **Sort Dialog Box :** For more advanced sorting options, you can use the Sort dialog box. To access it, go to the Data tab, click on the sort button, and choose the column(s) you want to sort by. In the dialog box, you can specify the sort order (ascending or descending) and add multiple levels of sorting.
3. **Custom Sort :** If you need to sort your data based on specific criteria, such as sorting by a custom list or sorting by cell color, you can use the Custom Sort features. This allows you to define your own sorting rules and apply them to your table.
4. **Sorting by Multiple Columns :** Excel also allows you to sort your table by multiple columns. This can be useful when you want to sort your data by one column and then by another column within the sorted results. To do this, use the Sort dialog box and add multiple levels of Sorting.

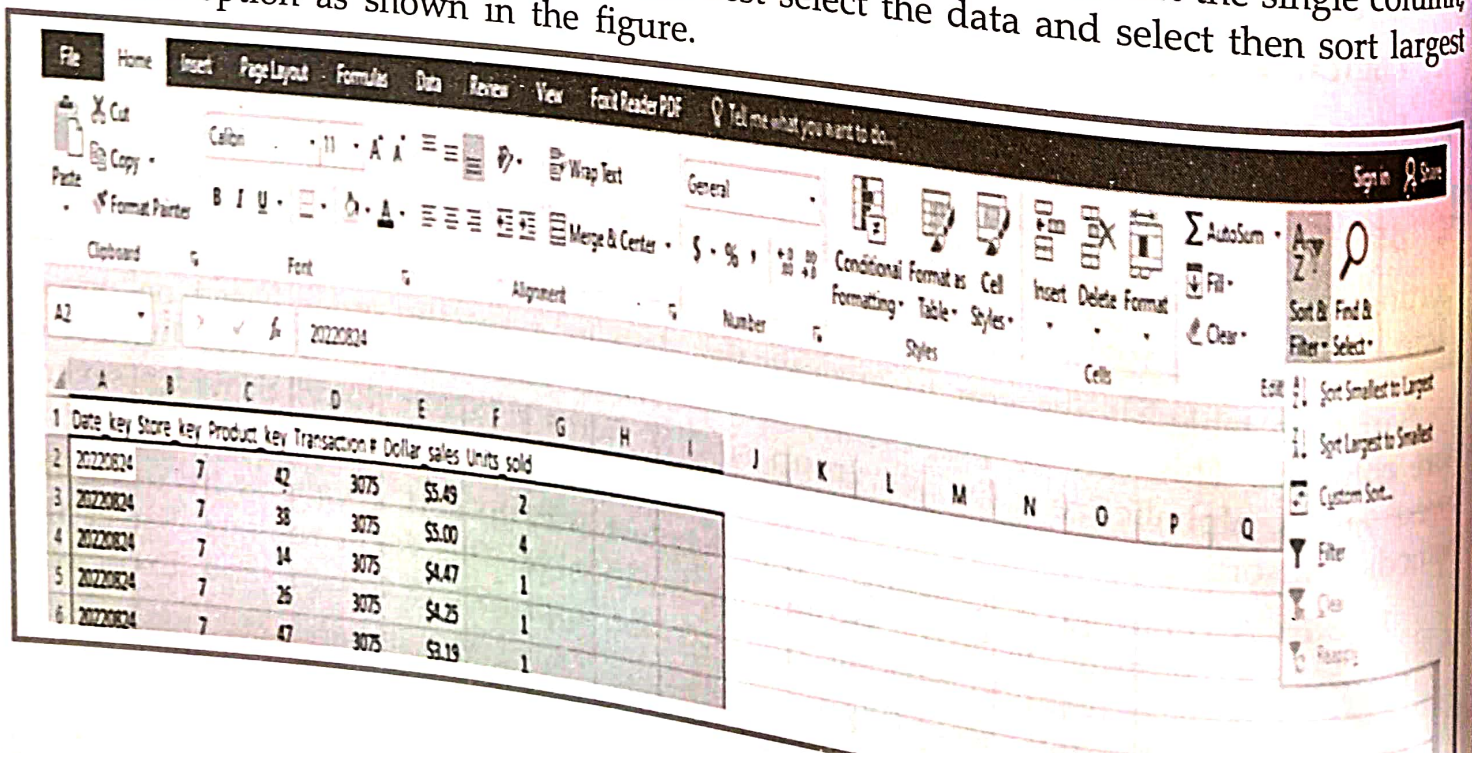
6.3.1 SORT ASCENDING/DESCENDING

When you need to sort the data list only on one particular field (such as the ID No, Last Name, or Location field), you simply click that field's AutoFilter button and then select the appropriate sort option from its drop-down list:

- Sort A to Z or Sort Z to A in a text field
- Sort Smallest to Largest or Sort Largest to Smallest in a number field
- Sort Oldest to Newest or Sort Newest to Oldest in a date field

When you choose an option, Excel immediately reorders the records, and then places a tiny arrow in the column header to indicate that you used this column for your sort. However, Excel doesn't keep re-sorting your data when you make changes or add new records (after all, it would be pretty distracting to have your records jump around unexpectedly). If you make changes and want to reapply the sort, go to the column header menu and choose the same sort option again.

Consider the sales data we have created in previous section. To sort the single column, Date_key in this case, to just largest to smallest select the data and select then sort largest to smallest option as shown in the figure.

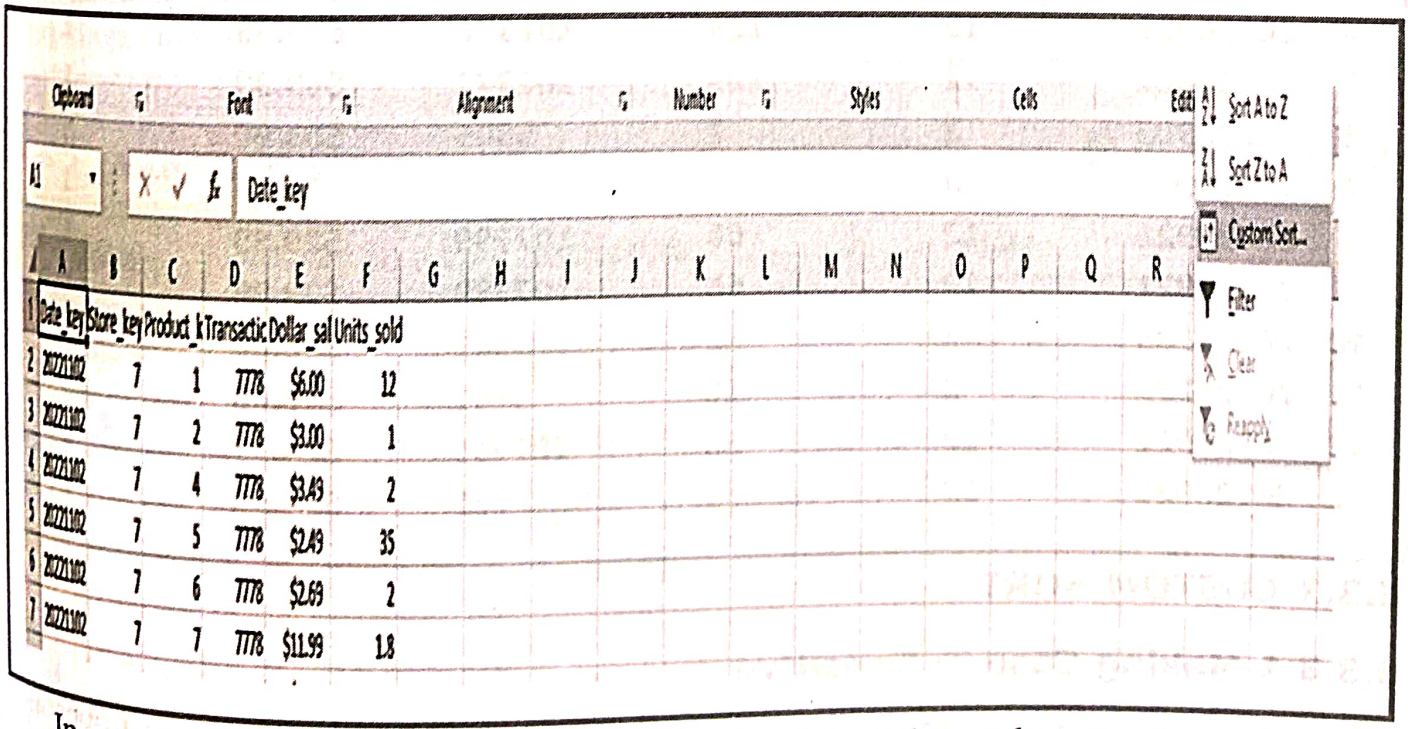


The resultant data is shown below:

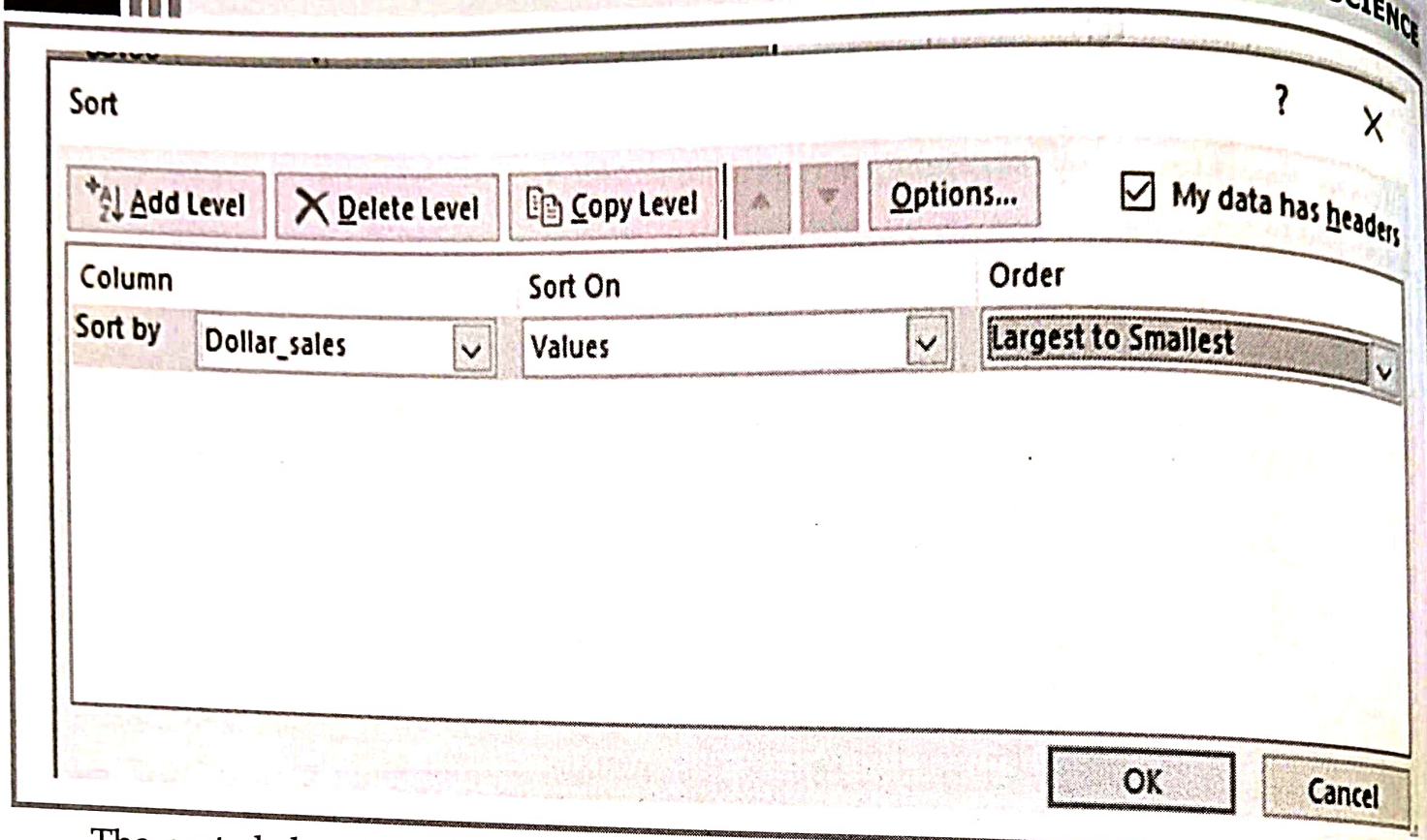
	A	B	C	D	E	F	G
1	Date_key	Store_key	Product_key	Transaction #	Dollar_sales	Units_sold	
2	20221111	2	48	4356	$8.68	1	
3	20221111	2	110	4356	$5.97	4	
4	20221111	2	43	4356	$5.58	3	
5	20221111	2	23	4356	$4.98	1	
6	20221111	2	14	4356	$4.47	1	
7	20221111	2	21	4356	$4.47	1	
8	20221111	2	26	4356	$4.25	1	
9	20221111	2	26	4356	$3.58	4	
10	20221111	2	49	4356	$3.48	4	
11	20221111	2	60	4356	$3.48	2	
12	20221111	2	23	4356	$3.38	1	
13	20221111	2	43	4356	$3.33	3	
14	20221111	2	26	4356	$3.33	3	

6.3.2 SORT DIALOG BOX

The Sort dialog in Excel offers up to 64 different sorting levels. In Sort dialog box, in which you can pick any number of columns to use as sort criteria andchoose whether to sort the rows in ascending or descending order. Using Sort dialog you can sort specific column. For example to sort the dollar_sale from largest to smallest select all the data and then select sort and filter and then custom sort tab using home tab.



In sort dialog box select 'my data has headers' and then select sort by option and select Dollar_sales. Select Order as 'Largest to Smallest'



The sorted data is as follows:

	A	B	C	D	E	F
1	Date_key	Store_key	Product_key	Transaction #	Dollar_sales	Units_sold
2	20220921	12	129	101372	$99.99	1
3	20220921	12	129	101372	$99.99	1
4	20220921	12	129	101372	$89.99	1
5	20220921	12	129	101372	$89.99	1
6	20220906	2	2	5809	$53.90	5
7	20220906	2	25	5809	$31.47	1
8	20220921	12	66	107799	$29.99	1
9	20220921	12	66	107799	$29.99	1
10	20221105	2	10	3941	$26.00	10
11	20221105	2	10	3941	$26.00	10
12	20220921	12	66	107799	$24.99	1
13	20220921	10	9	5282285	$24.99	1

6.3.3 CUSTOM SORT

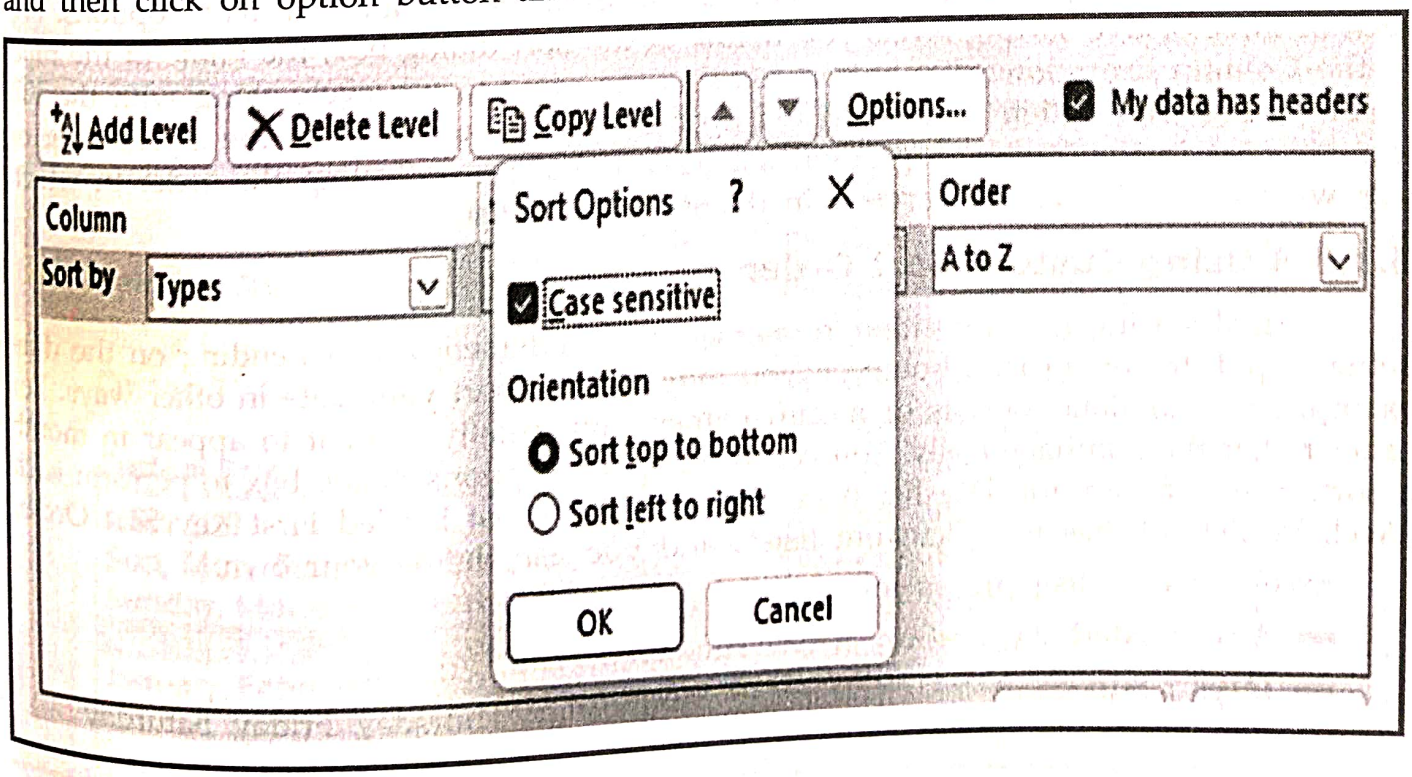
6.3.3.1 Making Case in Sensitive

Typically, an Excel sort ignores the case of the text. Values that are lowercase, uppercase or any combination of the two are treated equally in a sort. In Excel, a case-sensitive sort

sorts lowercase values before uppercase values. For example, abc will sort before ABC. Similarly, ABc will sort before ABC.If you want Excel to consider case when sorting, you use the Options button on the Sort dialog. Follow is the sample food sales database which has 'Types' column. It contains the type values in both lowercase and uppercase.

	A	B	C	D
1	Month	Types	Category	Amount
2	Jan	Personal	Restaurant	35
3	Jan	Business	Taxi	42
4	Jan	business	Restaurant	64
5	Jan	Personal	Restaurant	22
6	Jan	Business	Tips	12
7	Feb	Personal	Restaurant	32
8	Feb	personal	Entertainme	24
9	Feb	business	Entertainme	45
10	Feb	Business	Lodging	98
11	Mar	Business	Airfare	250
12	Mar	personal	Restaurant	24
13	Mar	Business	Restaurant	45
14	Mar	Business	Lodging	160
15	Mar	Business	Tolls	8

Click on custom sort options then Sort dialog box will appears. Select 'Types column and then click on option button and select 'Case sensitive' check box.



Following figure shows the sorted 'Type' column data where lowercase values come first following by uppercase values.

1	Month	Types	Category	Amount
2	Jan	business	Restaurant	64
3	Feb	business	Entertainme	45
4	Jan	Business	Taxi	42
5	Jan	Business	Tips	12
6	Feb	Business	Lodging	98
7	Mar	Business	Airfare	250
8	Mar	Business	Restaurant	45
9	Mar	Business	Lodging	160
10	Mar	Business	Tolls	8
11	Feb	personal	Entertainme	24
12	Mar	personal	Restaurant	24
13	Jan	Personal	Restaurant	35
14	Jan	Personal	Restaurant	22
15	Feb	Personal	Restaurant	32

6.3.3.2 Sort by Color Icon

Excel can sort data by fill color, font color, or icon sets. This also works with color applied through conditional formatting or color that you applied by using the cell format icons. Because color is subjective, there is not a default color sequence. If one column contains 17 different colors, you will have to set up 17 rules in the Sort dialog just to sort by that one column.

You can sort the records in the data list order by all the fill colors or cell icons assigned by applying the Conditional Formatting Color Scales and Cell Icons options. For each of three or five sorting levels you define in the Sort dialog box, the name of the field in the Column drop-down list button remains the same in all levels along with the Fill Color or Cell Icon option in the Sort On drop-down list button. Only the actual color or icon selected in the Order drop-down list button changes, reflecting the order in which you want to see the records appear in the sorted data list.

6.3.3.3 Using Custom Sort Order

Normal sorting is done either numerically or alphabetically, depending on the data being sorted. In some cases, however, you may want to sort your data in other ways. For example, if your data consists of month names, you usually want it to appear in month order rather than alphabetically. You can use the Sort Options dialog box to perform such a sort. Select the appropriate list from the drop-down list labeled First Key Sort Order, Excel, by default, has four "custom lists", and you can define your own.

Excel's custom lists are as follows :

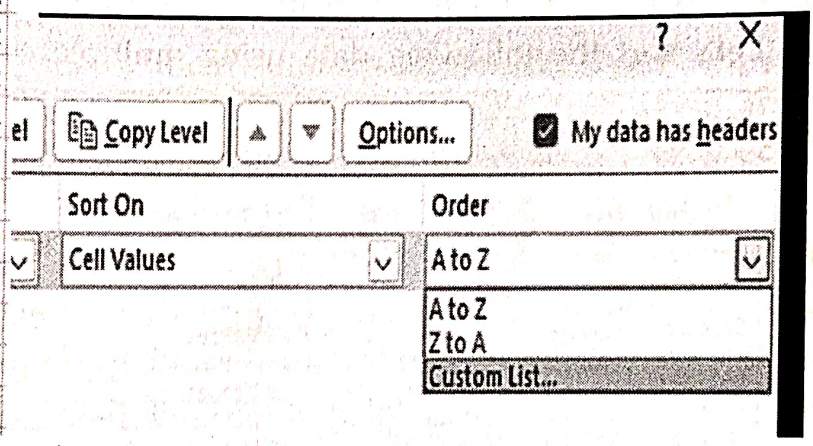
- **Abbreviated days :** Sun, Mon, Tue, Wed, Thu, Fri, Sat
- **Days :** Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday

- **Abbreviated Months :** Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec
- **Months :** January, February, March, April, May, June, July, August, September, October, November, December.

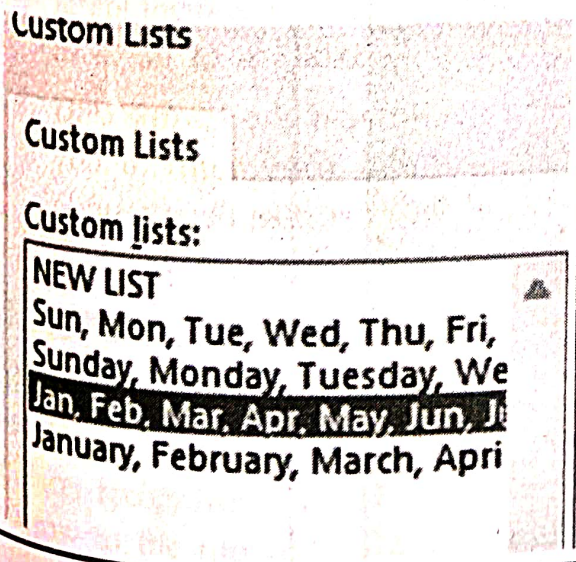
Note that the abbreviated days and months do not have periods after them. If you use periods for these abbreviations, they are not recognized (and are not sorted correctly).

For example, in food sales data. Sales is done monthly as shown below. To sort the data month wise select column 'month' and then select 'custom list' from order drop down list in the custom sort dialog box.

	Month	Types	Category	Amount
2	Jan	Personal	Restaurant	35
3	Nov	Business	Taxi	42
4	Jan	Business	Restaurant	64
5	Sep	Personal	Restaurant	22
6	Jan	Business	Tips	12
7	Feb	Personal	Restaurant	32
8	Oct	Personal	Entertainme	24
9	Feb	Business	Entertainme	45
10	Jul	Business	Lodging	98
11	Apr	Business	Airfare	250
12	Mar	Personal	Restaurant	24
13	May	Business	Restaurant	45
14	Mar	Business	Lodging	160
15	Jun	Business	Tolls	8



A custom list dialog box will appear. Select abbreviated months from this list. The sorted list is as follows:



Month	Types	Category	Amount
Jan	Personal	Restaurant	35
Jan	Business	Restaurant	64
Jan	Business	Tips	12
Feb	Personal	Restaurant	32
Feb	Business	Entertainme	45
Mar	Personal	Restaurant	24
Mar	Business	Lodging	160
Apr	Business	Airfare	250
May	Business	Restaurant	45
Jun	Business	Tolls	8
Jul	Business	Lodging	98
Sep	Personal	Restaurant	22
Oct	Personal	Entertainme	24
Nov	Business	Taxi	42

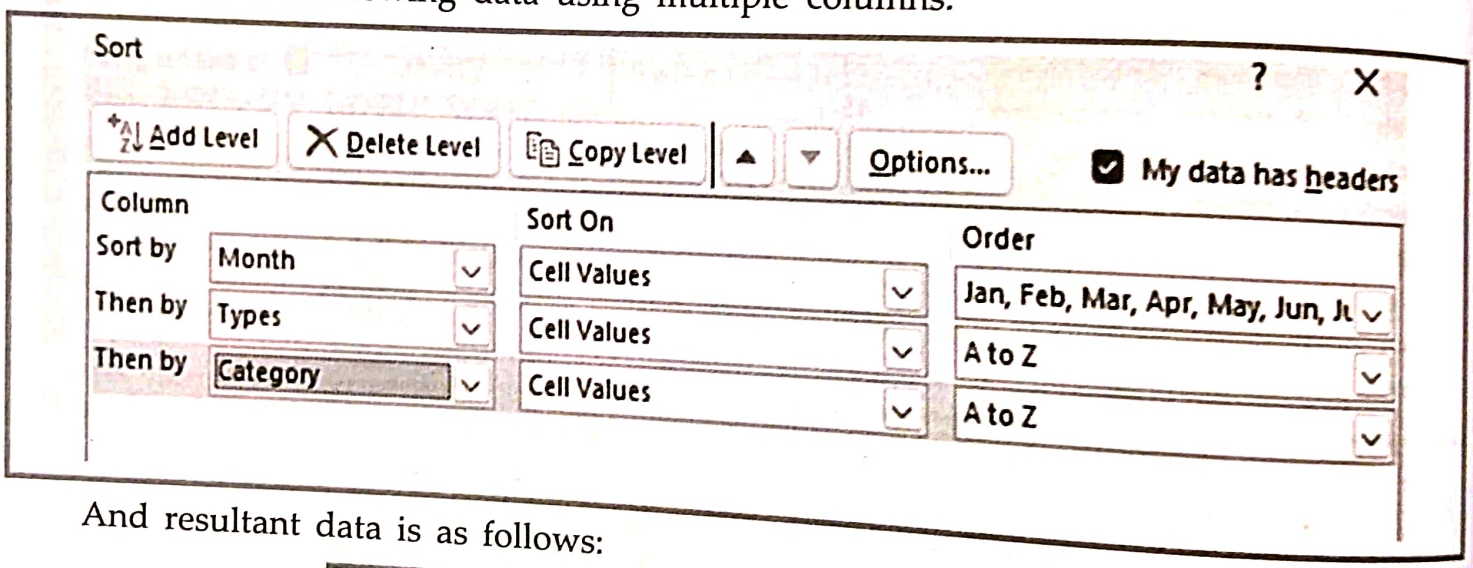
6.3.4 SORTING BY MULTIPLE COLUMNS

When you need to sort a data list on more than one field, you use the Sort dialog box. And you need to sort on more than one field when the first field contains duplicate values and you want to determine how the records with duplicates are arranged. When you need to sort a data list on more than one field, you use the Sort dialog box. And you need to sort on more than one field when the first field contains duplicate values and you want to determine how the records with duplicates are arranged.

Sometimes, you may need to sort on a whole bunch of fields to get the desired order. For example, suppose in food sales data you want to find monthly sales based on it type and category. Soring of multiple columns will be done as follows

- First sort by Month field using months abbreviation (as we did in the last section).
- Then by Type field in A to Z order
- Then by Categpru field in A to Z order

Let's sort the following data using multiple columns.



And resultant data is as follows:

Month	Types	Category	Amount
Jan	Business	Restaurant	64
Jan	Business	Tips	12
Jan	Personal	Restaurant	35
Feb	Business	Entertainme	45
Feb	Personal	Restaurant	32
Mar	Business	Lodging	160
Mar	Personal	Restaurant	24
Apr	Business	Airfare	250
May	Business	Restaurant	45
Jun	Business	Tolls	8
Jul	Business	Lodging	98
Sep	Personal	Restaurant	22
Oct	Personal	Entertainme	24
Nov	Business	Taxi	42

6.4 FILTERING DATA

One of the biggest problems with large tables is that it's often hard to find and extract the data you need. Sorting can help, but in the end, you're still working with the entire table. You need a way to define the data you want to work with and then have Excel display only those records on screen. This is called **filtering your** data, and Excel's Filter command makes filtering out subsets of your data as easy as selecting an option from a drop-down list.

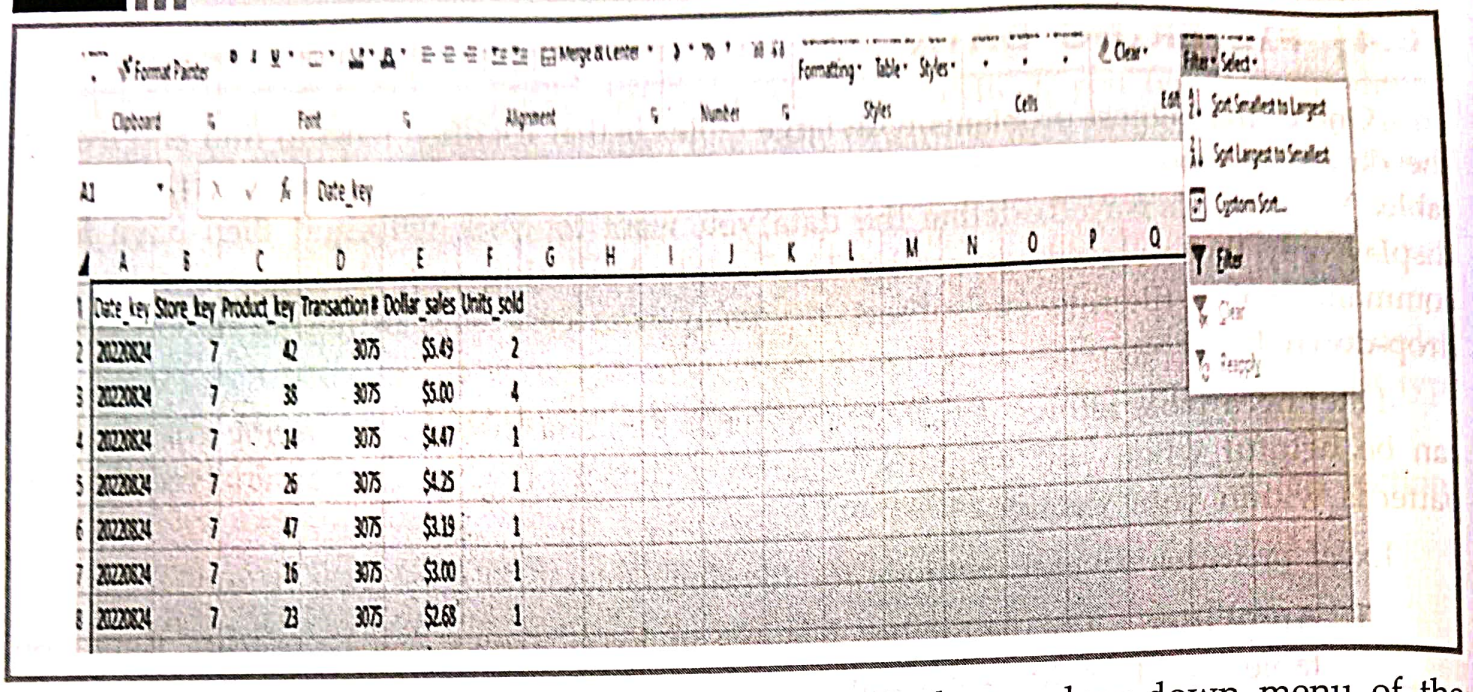
Filtering a table allows you to display only the data that meets specific criteria. This can be helpful when you want to focus on a subset of your data or analyze specific patterns within your dataset.

Excel provides several methods for filtering tables :

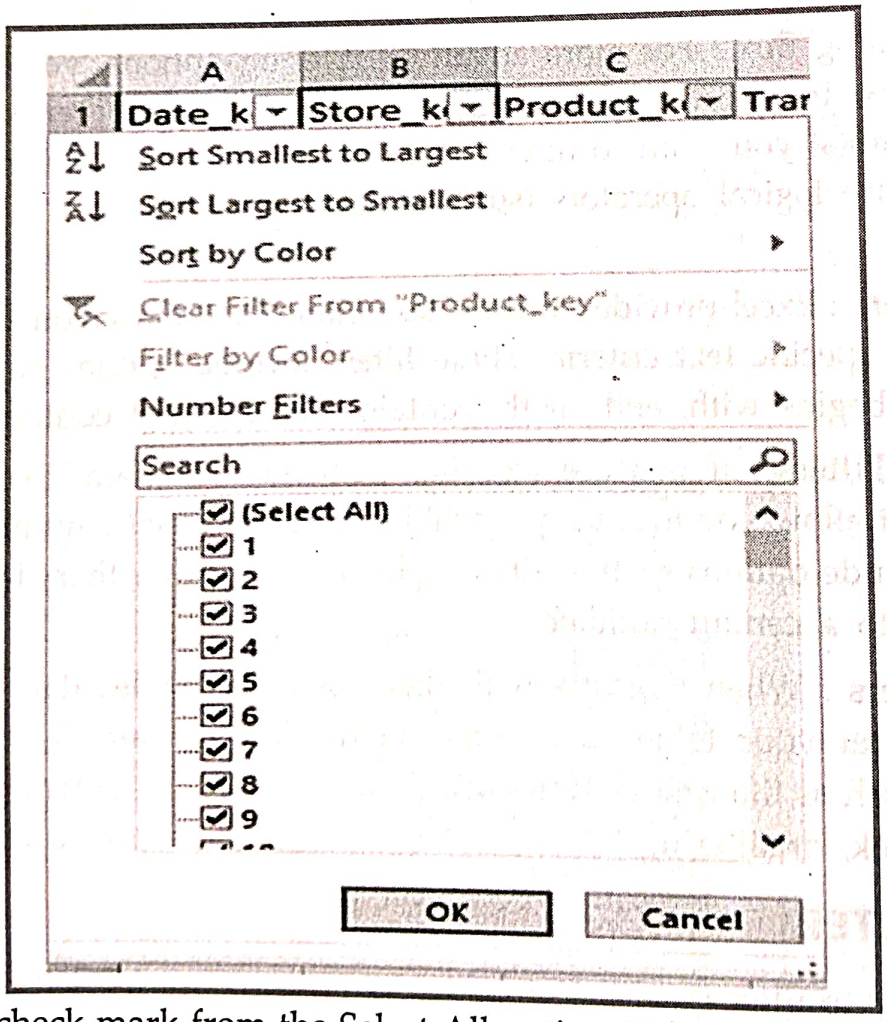
1. **AutoFilter :** The AutoFilter features in Excel allows you to quickly filter you table based on the values in one or more columns. To enable AutoFilter, go to the Data tab, click on the Filter button, and Excel will add drop-down arrows to each column header. You can then use these drop-down arrows to select the values you want to display.
2. **Filter Dialog Box :** For more advanced filtering options, you can use the filter dialog box. To access it, go to the Data tab, click on the Filter button, and choose the column(s) you want to filter by, In the dialog box, you can specify multiple criteria, use logical operators (such as AND and OR), and even create custom filters.
3. **Text Filters :** Excel provides various text filters that allow you to filter your table based on specific text criteria. These filters include options such as filtering by text that begins with, ends with, contains, or does not contain a certain value.
4. **Number Filters :** If you are working with numeric data, Excel offers number filters that allow you to filter your table based on specific number criteria. These filters include options such as filtering by values greater than, less than, between, or equal to a certain number.
5. **Data Filters :** When working with dates, Excel provides date filters that allow you to filter your table based on specific date criteria/ These filters include options such as filtering by dates within a certain range, filtering by specific days of the week, or filtering by dates that are before or after a certain date.

6.4.1 AUTO FILTER

The easiest way to filter your data is to use the AutoFilter feature in Excel. To create an AutoFilter, click anywhere in your worksheet table and choose Data > Sort & Filter > Filter. Excel recognizes that all of the adjacent cells are part of your table and applies a filter arrow to the top cell in each column, as seen in Figure



Click the filter arrow in each column header displays a drop-down menu of the unique entries in that column. In Figure, you can see that the unique values in the product_key.



Remove the check mark from the Select All option. The check marks are automatically removed from the rest of the unique entries. Click the filtering option you want to see on your worksheet and click OK to apply the filter.

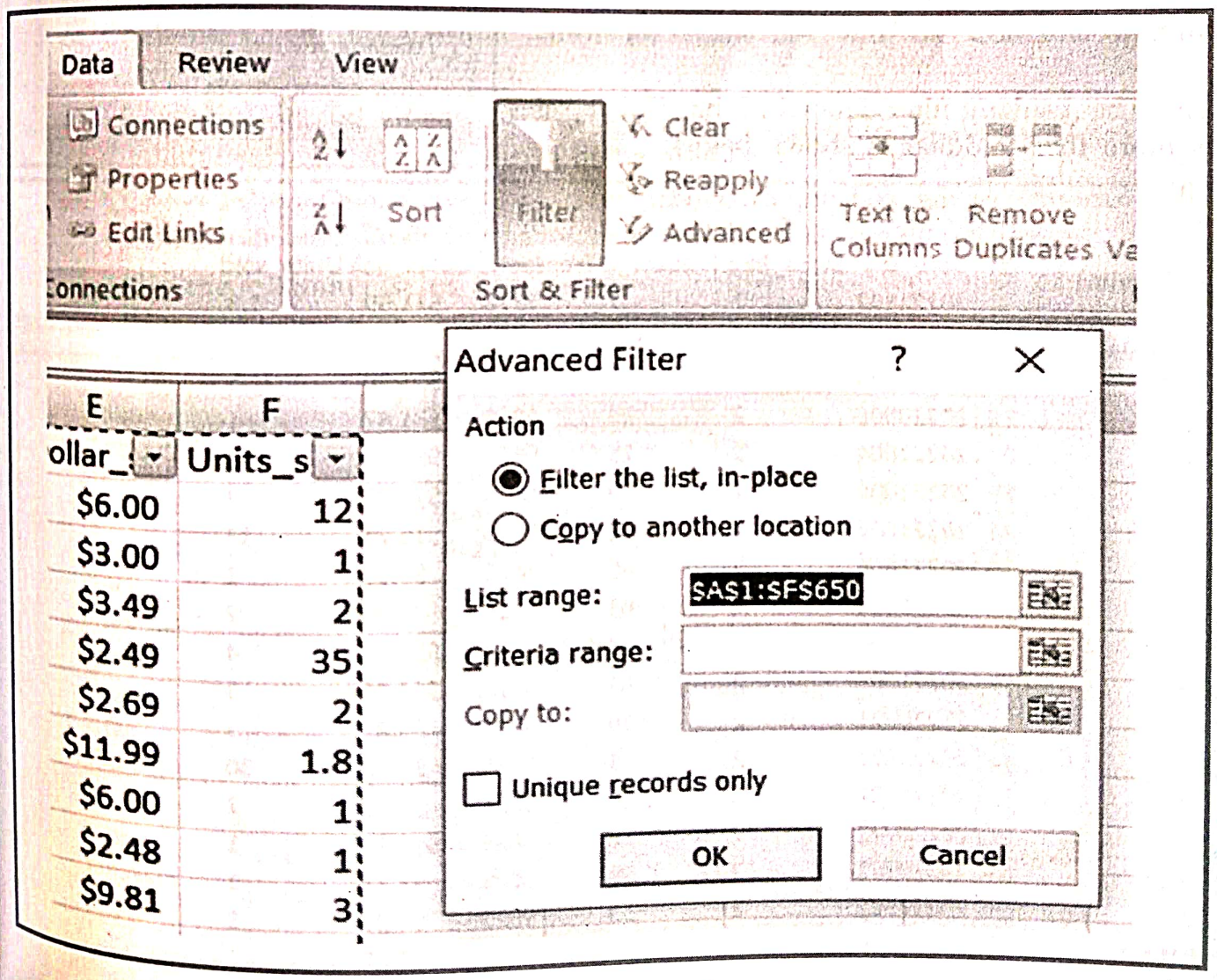
To remove the filters from the worksheet and return to the full unfiltered list of data, choose one of the following options:

- Click the filter button next to the column header and choose Clear Filter.
- Click the Select All option and then click OK.
- Choose Data > Filter from the Ribbon.

6.4.2 FILTER DIALOG BOX

The Advanced Filter command, unlike the standard filter, requires that you specify filtering criteria in a worksheet range separate from your list or table. Because entire rows are hidden when the filter is executed, it's inadvisable to put the criteria range *alongside* the list. Instead, put it above the list or on a separate worksheet. A criteria range must consist of at least two rows. Type one or more column headings in the top row, and type your filtering criteria in the second and subsequent rows. With the exception of computed criteria, you must spell the headings in your criteria range exactly like those in your list.

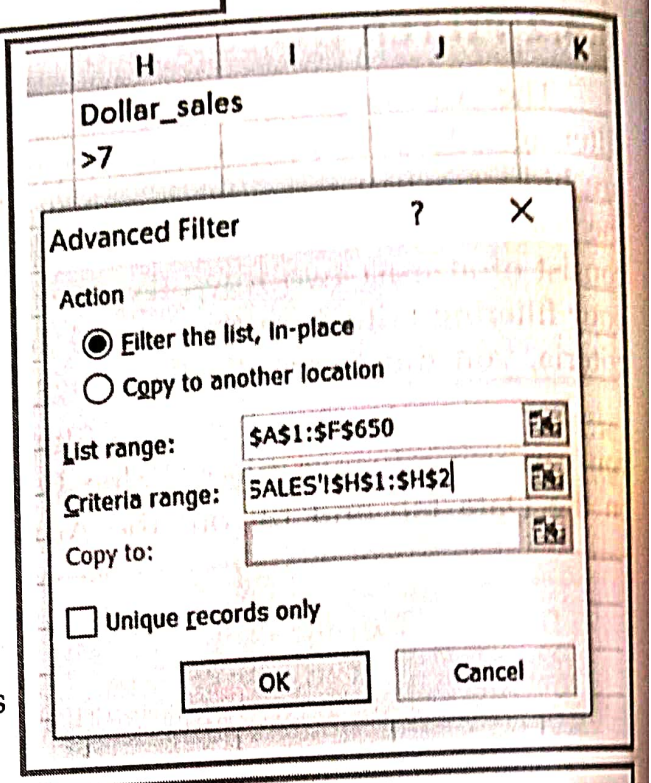
For example, if you want to filter the rows whose Dollar_sale more than 7 dollar then you can use advance filter dialog box. Click Advanced Filter in the Sort & Filter group on the Data tab and fill out the Advanced Filter dialog box as shown in Figure



Now write criteria in the workbook show below:

	F	G	H
	Units_s		Dollar_sales
	12		>7
	1		

Add the criteria range in the advance filter. Next you have 2 options. First options is for if you apply the filter and show the result by hiding the original data rows. Second options copy the result to the specified location. Select first option as shown below:



The resultant filter data which has Dollar_sales is more than 7 dollar is shown below:

	A	B	C	D	E	F	G
1	Date_key	Store_key	Product_k	Transactio	Dollar_sal	Units_sold	
7	20221102	7	7	7778	$11.99	1.8	
10	20220906	2	19	5809	$9.81	3	
19	20220906	2	2	5809	$53.90	5	
20	20220906	2	25	5809	$31.47	1	
27	20221004	2	15	5625	$10.62	6	
29	20221004	2	58	5625	$10.41	3	
34	20221004	2	23	5625	$15.21	13	
36	20221004	2	80	5625	$9.00	1	
37	20221004	2	81	5625	$21.96	2	
38	20221004	2	2	5625	$22.76	4	
45	20221004	2	84	5625	$14.01	3	
51	20220921	2	19	5625	$9.81	3	
52	20220921	2	1	5625	$8.48	30	
56	20220921	2	91	5625	$7.48	2	
57	20220921	2	88	5625	$10.97	1	
63	20220921	2	96	5625	$8.17	1	
64	20220921	2	53	5625	$7.07	1	

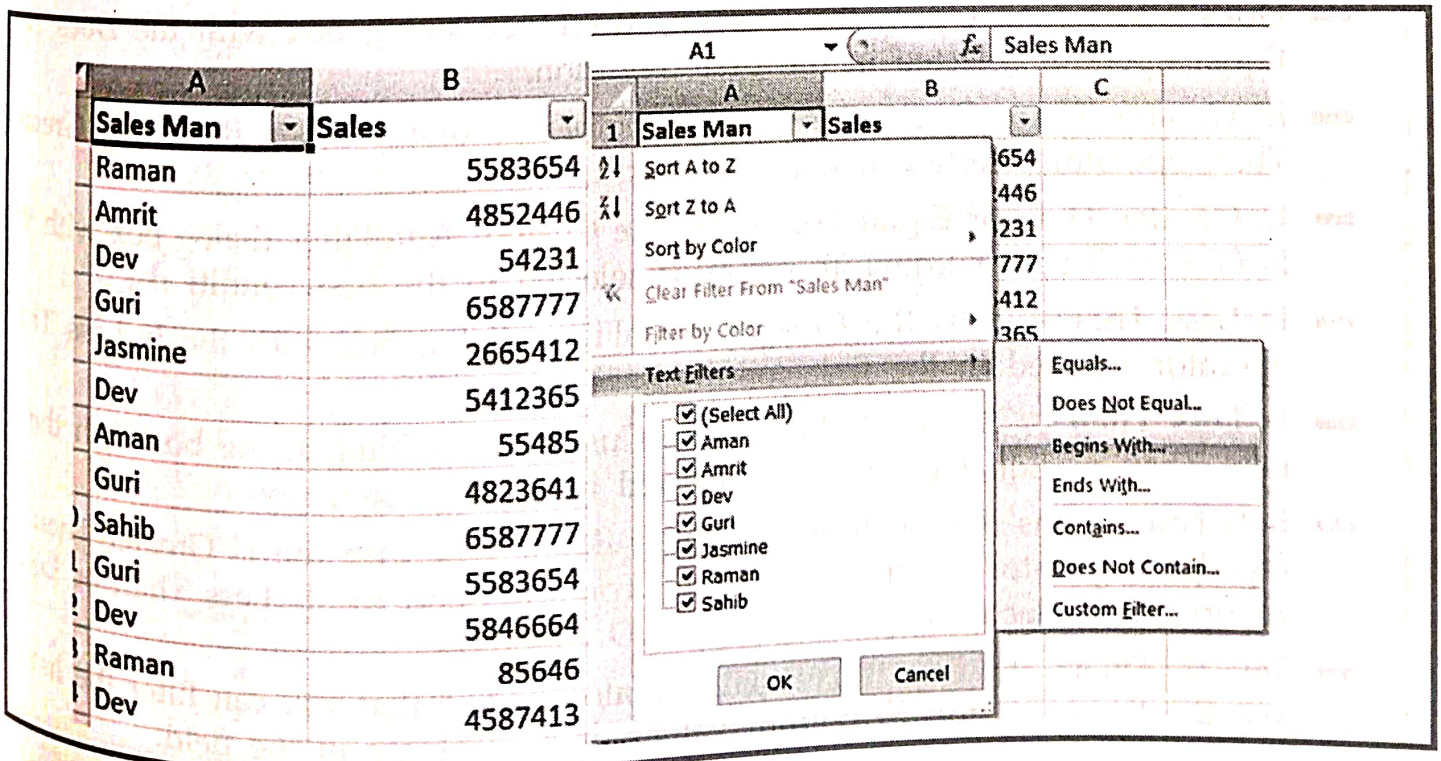
6.4.3 TEXT FILTER

For text, you can filter values that match exactly, or values that contain a piece of text. To apply text filtering, open the drop-down column list, and then choose Text Filters. When you filter text, you can get even more versatile results using wildcards. The asterisk (*) matches any series of characters, while the question mark (?) matches a single character. So the filter expression Category equals T* matches any category that starts with the letter T. The filter expression Category equals T???? matches any five-letter category that starts with T.

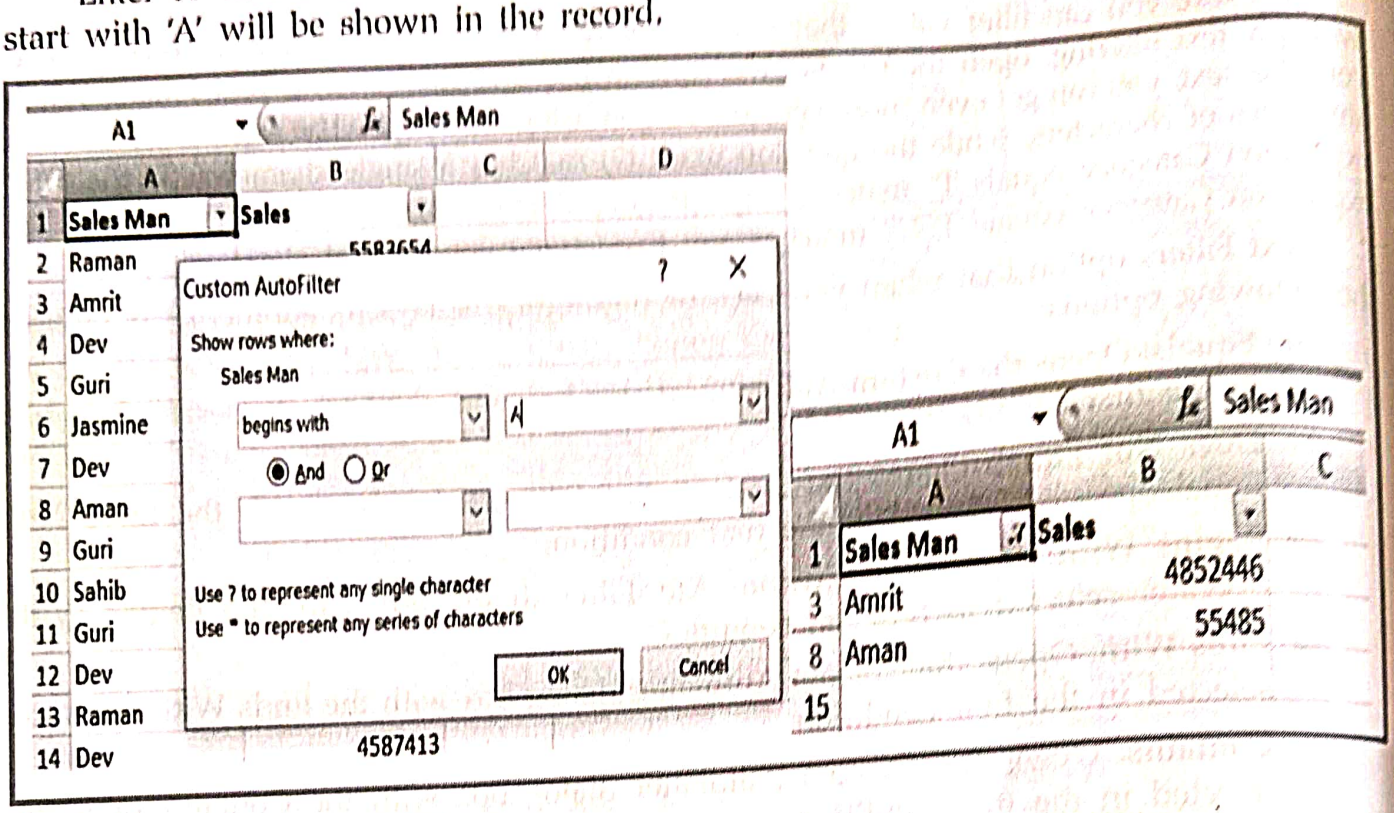
Text Filters option that when you click or highlight displays its submenu containing the following options:

- **Equals:** Opens the Custom AutoFilter dialog box with the Equals operator selected in the first condition.
- **Does Not Equal:** Opens the Custom AutoFilter dialog box with the Does Not Equal operator selected in the first condition.
- **Begins With:** Opens the Custom AutoFilter dialog box with the Begins With operator selected in the first condition.
- **Ends With:** Opens the Custom AutoFilter dialog box with the Ends With operator selected in the first condition.
- **Contains:** Opens the Custom AutoFilter dialog box with the Contains operator selected in the first condition.
- **Does Not Contain:** Opens the Custom AutoFilter dialog box with the Does Not Contain operator selected in the first condition.
- **Custom Filter:** Opens the Custom AutoFilter dialog box where you can select your own criteria for applying more complex AND or OR conditions.

Consider the customer sales data as shown below: Click the arrow of Sales_Man field and then select text filter and then select Beging with option.



Enter 'A' in the Custom Autofilter and click ok. Only those Sales_Man whose name start with 'A' will be shown in the record.



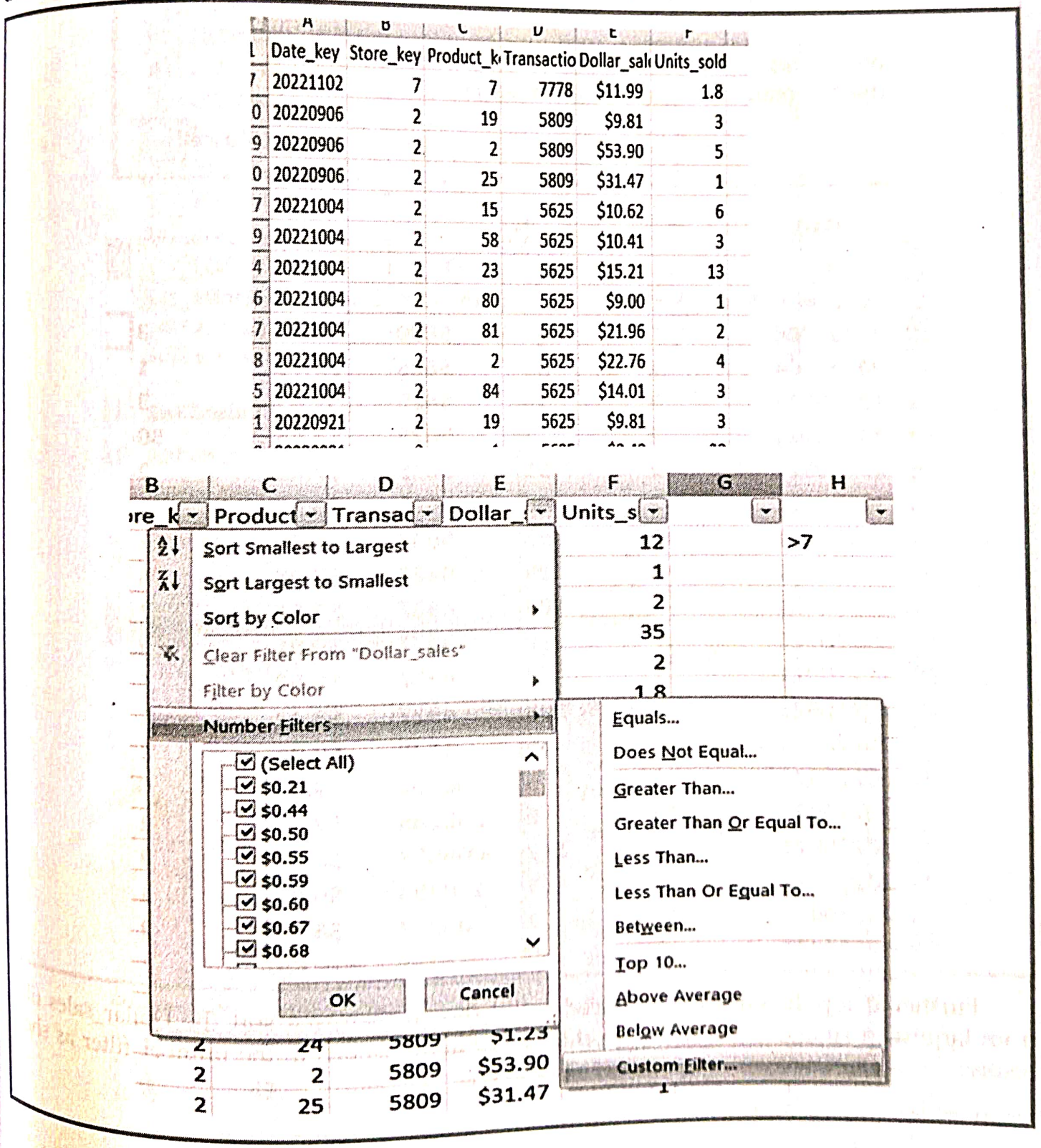
6.4.4 NUMBER FILTER

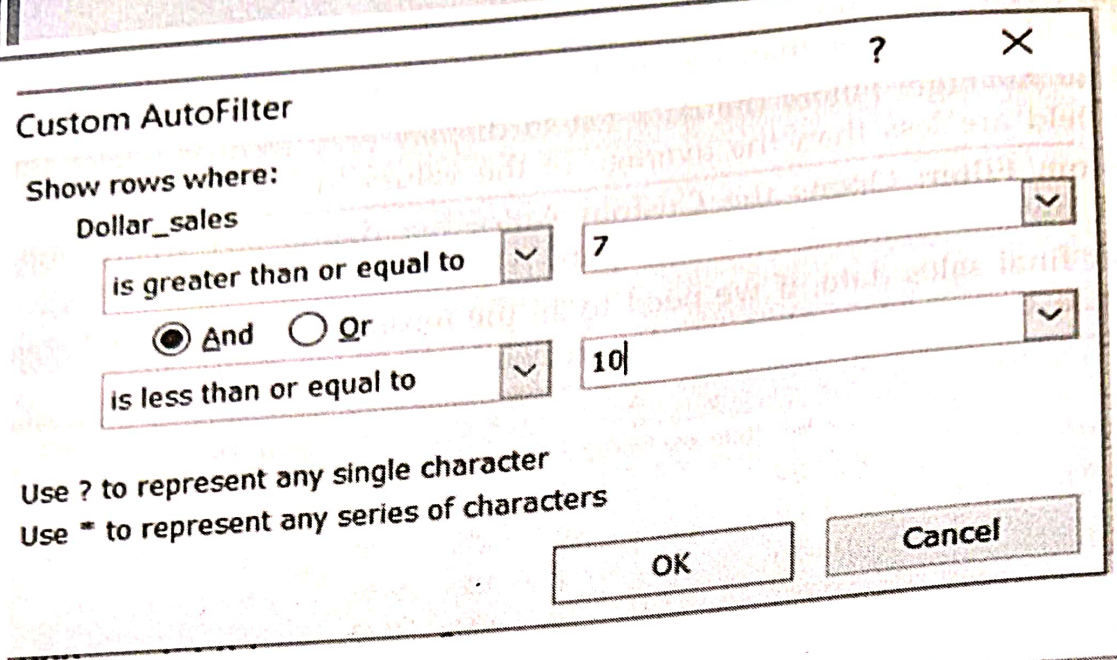
The AutoFilter drop-down menu for a field that contains only number entries besides dates or a combination of dates and other numeric entries contains a Number Filters option that when you click or highlight it displays its submenu containing the following options:

- **Equals:** Opens the Custom AutoFilter dialog box with the Equals operator selected in the first condition.
- **Does Not Equal:** Opens the Custom AutoFilter dialog box with the Does Not Equal operator selected in the first condition.
- **Is Greater Than:** Opens the Custom AutoFilter dialog box with the Is Greater Than operator selected in the first condition.
- **Is Greater Than or Equal To:** Opens the Custom AutoFilter dialog box with the Is Greater Than or Equal To operator selected in the first condition.
- **Is Less Than:** Opens the Custom AutoFilter dialog box with the Is Less Than operator selected in the first condition.
- **Is Less Than or Equal To:** Opens the Custom AutoFilter dialog box with the Is Less Than or Equal to operator selected in the first condition.
- **Between:** Opens the Custom AutoFilter dialog box with the Is Greater Than or Equal To operator selected in the first condition and the Is Less Than or Equal To operator selected in the second AND condition.
- **Top 10:** Opens the Top 10 AutoFilter dialog box so that you can filter the list to just the ten or so top or bottom values or percentages in the field.

- **Above Average:** Filters the data list to display only records where the values in the field are greater than the average of the values in this field.
- **Below Average:** Filters the data list to display only records where the values in the field are less than the average of the values in this field.
- **Custom Filter:** Opens the Custom AutoFilter dialog box where you can select your own criteria for more complex AND or OR conditions.

Consider final sales data, if we need to all the records whose sales is between 7 to 10 dollar then use the number filter as shown below:



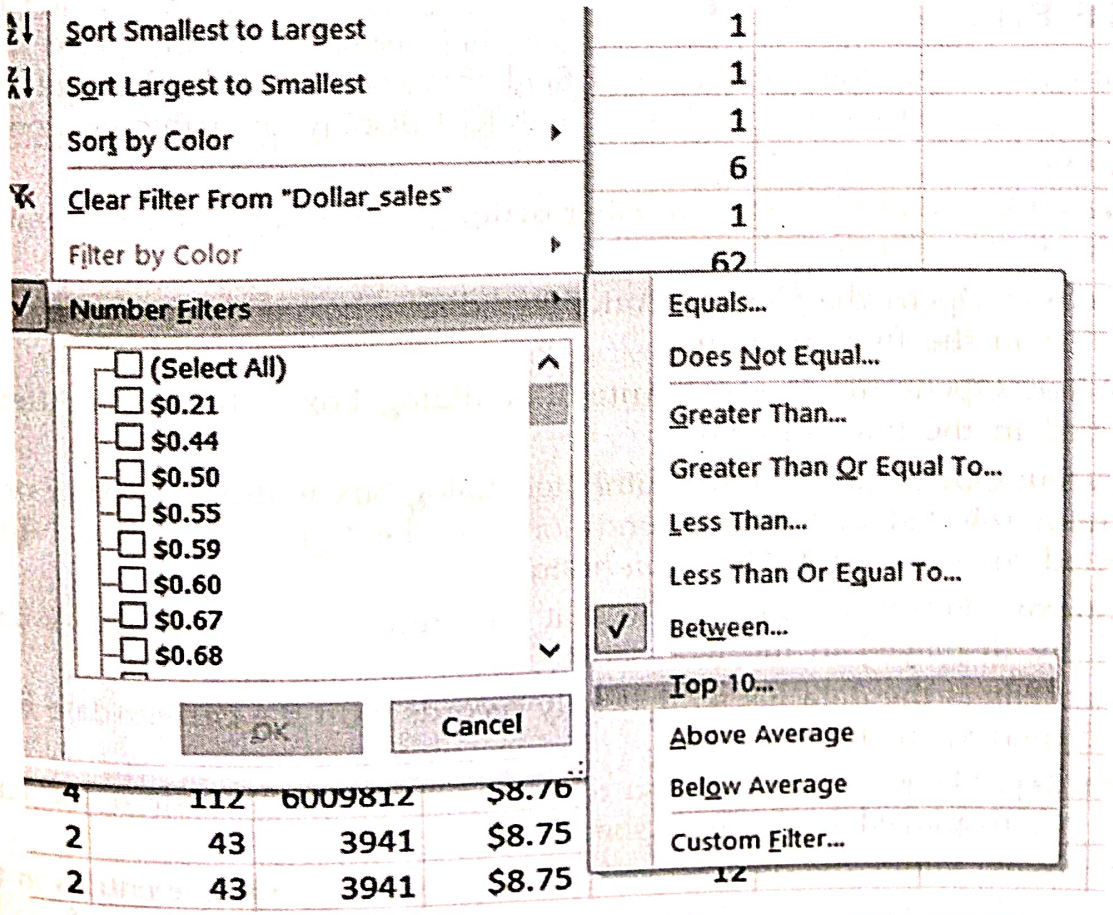


G10

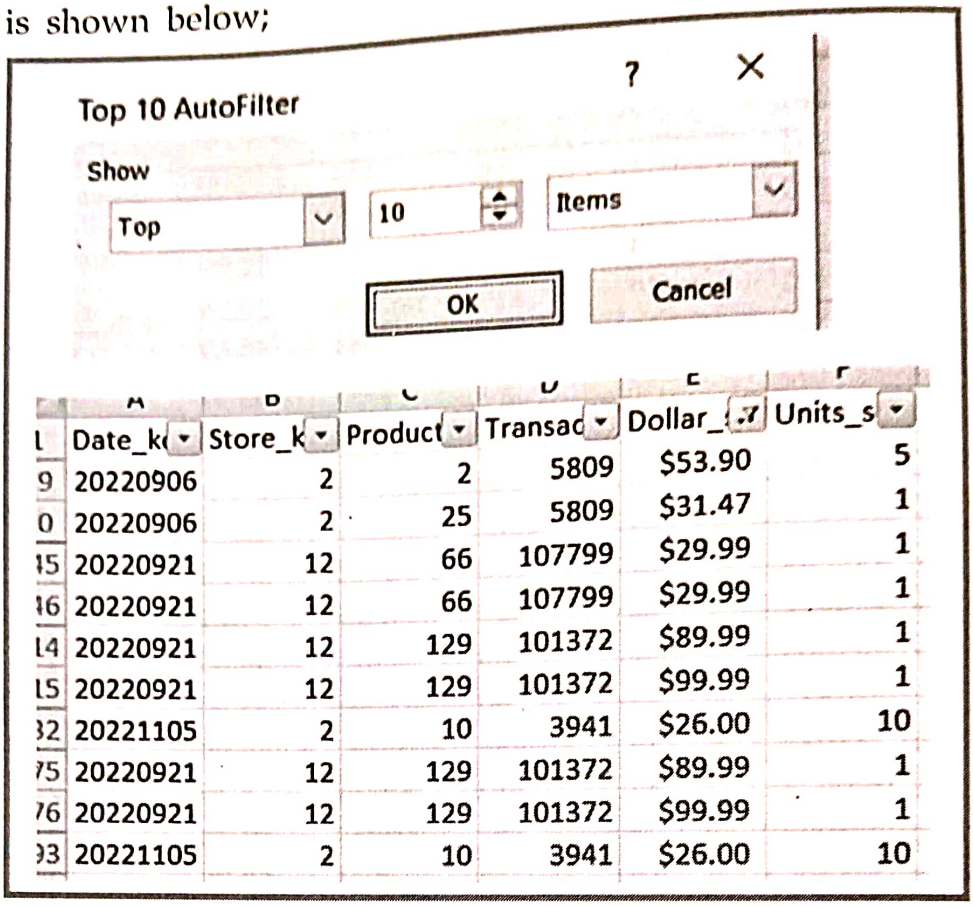
	A	B	C	D	E	F
1	Date_k	Store_k	Product	Transac	Dollar_	Units_s
10	20220906	2	19	5809	$9.81	3
36	20221004	2	80	5625	$9.00	1
51	20220921	2	19	5625	$9.81	3
52	20220921	2	1	5625	$8.48	30
56	20220921	2	91	5625	$7.48	2
53	20220921	2	96	5625	$8.17	1
54	20220921	2	53	5625	$7.07	1
55	20220921	2	15	4327	$7.08	4
31	20220921	2	100	4327	$7.56	1
35	20220921	2	19	4327	$9.81	3
38	20220921	2	80	4327	$9.00	1
96	20220921	2	108	4327	$8.47	1
17	20220921	9	65	1128	$9.99	1
87	20220921	1	3	400900	$9.99	6
13	20220921	10	6	5282285	$7.78	1
38	20220921	4	112	5004553	$8.16	2
44	20220921	14	114	100353	$9.99	1
59	20220921	11	112	88317	$8.76	2
93	20220921	2	26	1478	$10.00	1

Further if top 10 sales are required to be shown then first sort the Dollar_sales filed from largest to smallest and then use the top 10 item options in the number filter as show below:

	A	B	C	D	E	F
1	Date_k	Store_k	Product	Transac	Dollar_	Units_s
10	20220921	2	26	1478	$10.00	1
36	20220921	2	26	1478	$10.00	1
51	20220921	9	65	1128	$9.99	1
52	20220921	1	3	400900	$9.99	6
56	20220921	14	114	100353	$9.99	1
63	20221105	2	81	3941	$9.97	62
64	20221105	2	81	3941	$9.97	62
65	20221105	2	36	3941	$9.85	2
81	20221105	2	36	3941	$9.85	2
85	20220906	2	19	5809	$9.81	3
88	20220921	2	19	5625	$9.81	3
96	20220921	2	19	4327	$9.81	3
117	20221004	2	80	5625	$9.00	1
187	20220921	2	80	4327	$9.00	1
213	20221027	11	138	4256790	$8.99	50
238	20220921	11	112	88317	$8.76	2
244	20221028	4	112	6009812	$8.76	1
259	20221105	2	43	3941	$8.75	12



The result is shown below;



	A	B	C	D	E	F
1	Date_k	Store_k	Product	Transac	Dollar_	Units_s
9	20220906	2	2	5809	$53.90	5
10	20220906	2	25	5809	$31.47	1
15	20220921	12	66	107799	$29.99	1
16	20220921	12	66	107799	$29.99	1
14	20220921	12	129	101372	$89.99	1
15	20220921	12	129	101372	$99.99	1
32	20221105	2	10	3941	$26.00	10
75	20220921	12	129	101372	$89.99	1
76	20220921	12	129	101372	$99.99	1
93	20221105	2	10	3941	$26.00	10

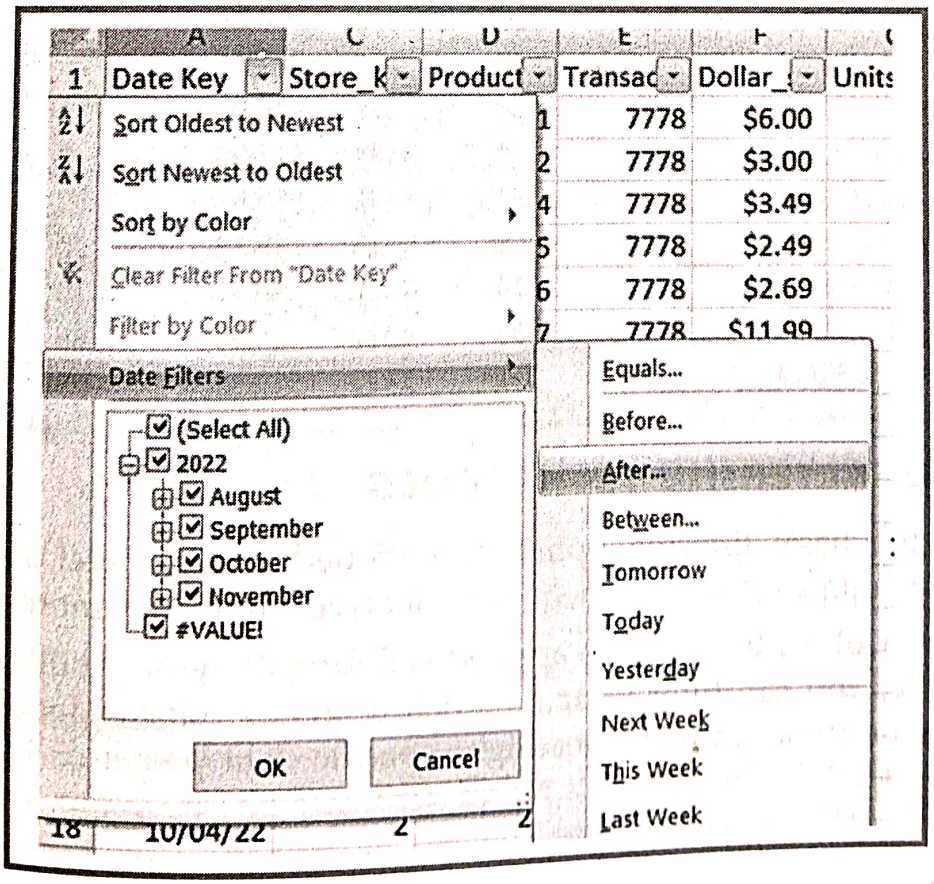
6.4.5 DATE FILTER

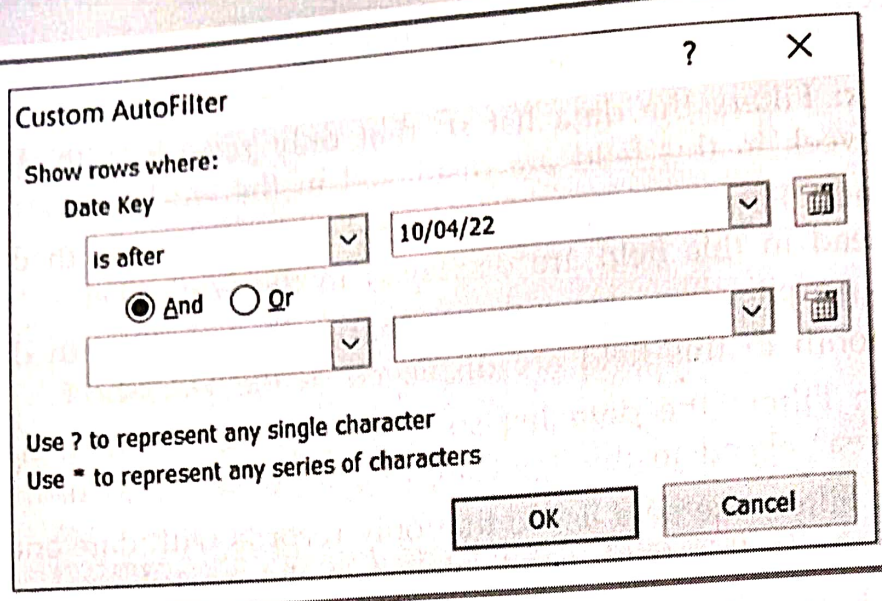
The AutoFilter drop-down menu for a field that contains only date entries contains a Date Filters option that when you click or highlight displays its submenu containing the following options:

- **Equals:** Opens the Custom AutoFilter dialog box with the Equals operator selected in the first condition.
- **Is Before:** Opens the Custom AutoFilter dialog box with the Is Before operator selected in the first condition.
- **Is After:** Opens the Custom AutoFilter dialog box with the Is After operator selected in the first condition.
- **Between:** Opens the Custom AutoFilter dialog box with the Is After or Equal To operator selected in the first condition and the Is Before or Equal To operator selected in the second AND condition.
- **Tomorrow:** Filters the data list so that only records with tomorrow's date in this field are displayed in the worksheet.
- **Today:** Filters the data list so that only records with the current date in this field are displayed in the worksheet.
- **Yesterday:** Filters the data list so that only records with yesterday's date in this field are displayed in the worksheet.
- **Next Week:** Filters the data list so that only records with date entries in the week ahead in this field are displayed in the worksheet.

- **This Week:** Filters the data list so that only records with date entries in the current week in this field are displayed in the worksheet.
- **Last Week:** Filters the data list so that only records with date entries in the previous week in this field are displayed in the worksheet.
- **Next Month:** Filters the data list so that only records with date entries in the month ahead in this field are displayed in the worksheet.
- **This Month:** Filters the data list so that only records with date entries in the current month in this field are displayed in the worksheet
- **Next Year:** Filters the data list so that only records with date entries in the calendar year ahead in this field are displayed in the worksheet.
- **This Year:** Filters the data list so that only records with date entries in the current calendar year in this field are displayed in the worksheet.
- **Last Year:** Filters the data list so that only records with date entries in the previous calendar year in this field are displayed in the worksheet.
- **Year to Date:** Filters the data list so that only records with date entries in the current year up to the current date in this field are displayed in the worksheet.
- **All Dates in the Period:** Filters the data list so that only records with date entries in the quarter (Quarter 1 through Quarter 4) or month (January through December) that you choose from its submenu are displayed in the worksheet.
- **Custom Filter:** Opens the Custom AutoFilter dialog box where you can select your own criteria for more complex AND or OR conditions.

Consider the final sales data, manager needs to get the sales detail after the date of 4 October 2022 then date filter does the job as shown below in the figures.





And resultant figure is as follows:

1	Date Key	Store_k	Product	Transac	Dollar_	Units_s
55	11/05/22	2	90	5625	$3.34	2
56	11/05/22	14	114	100353	$9.99	1
57	11/05/22	2	88	5625	$10.97	1
73	11/05/22	2	17	4327	$4.96	1
05	10/27/22	4	73	20001385	$2.48	1
36	10/28/22	4	73	5004553	$2.67	0.5
51	11/05/22	15	115	171813	$1.69	1
85	11/05/22	4	126	6009658	$0.99	1
04	11/11/22	2	77	1478	$2.97	90
15	10/30/22	12	129	101372	$99.99	1
16	10/30/22	2	87	667	$2.47	20
17	10/30/22	12	130	101372	$6.99	2
18	10/30/22	10	9	3144	$17.98	8
19	10/30/22	2	131	4322	$0.88	52
20	10/30/22	2	1	5625	$8.48	30
21	10/30/22	2	110	4322	$4.77	75
22	10/30/22	2	93	4322	$4.77	35
23	10/30/22	6	132	1427	$12.59	1.3

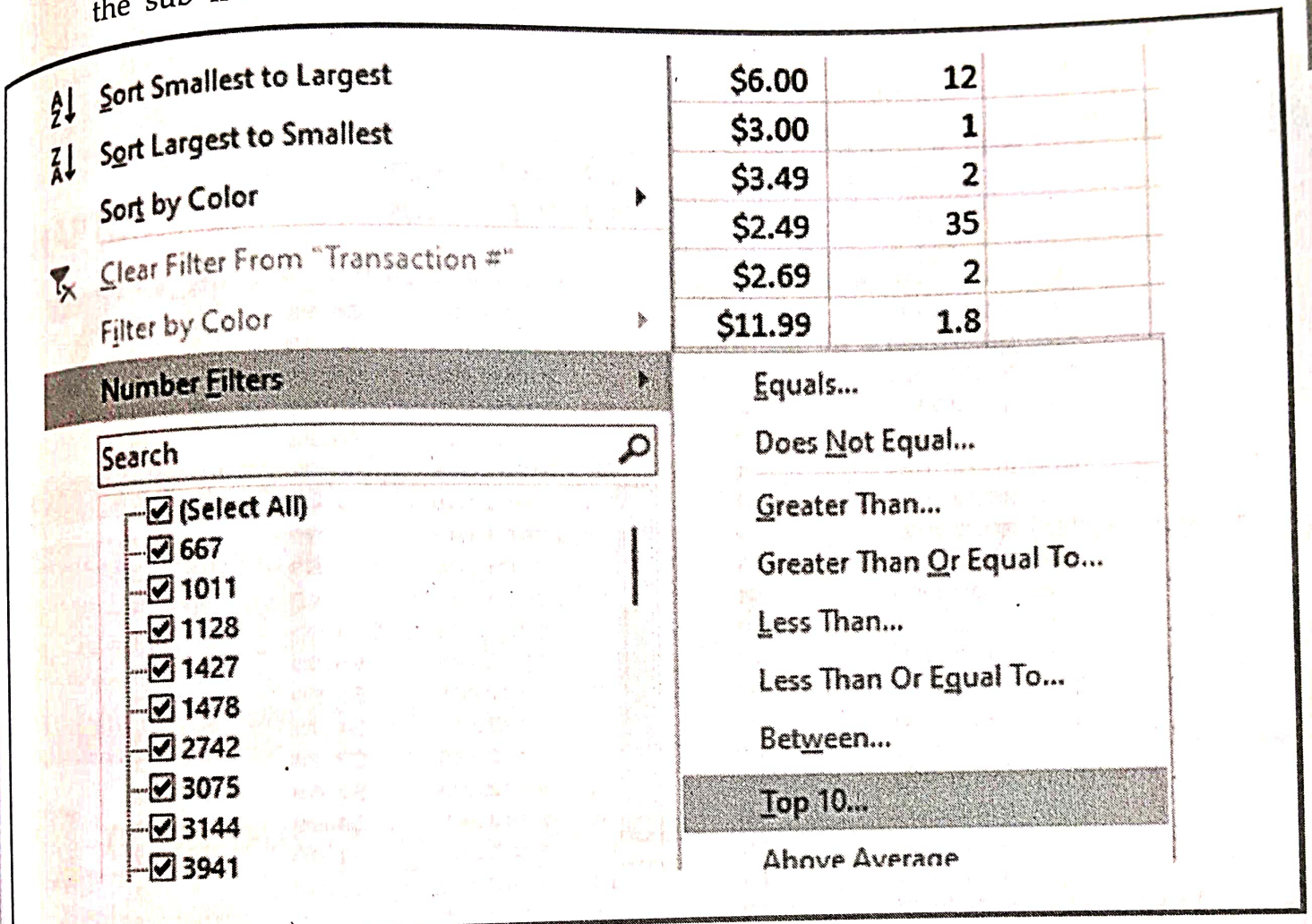
6.4.6 ADVANCED FILTERING TECHNIQUES

In addition to the basic filtering options mentioned above, Excel also offers some advanced filtering techniques that can help you analyze your data more effectively:

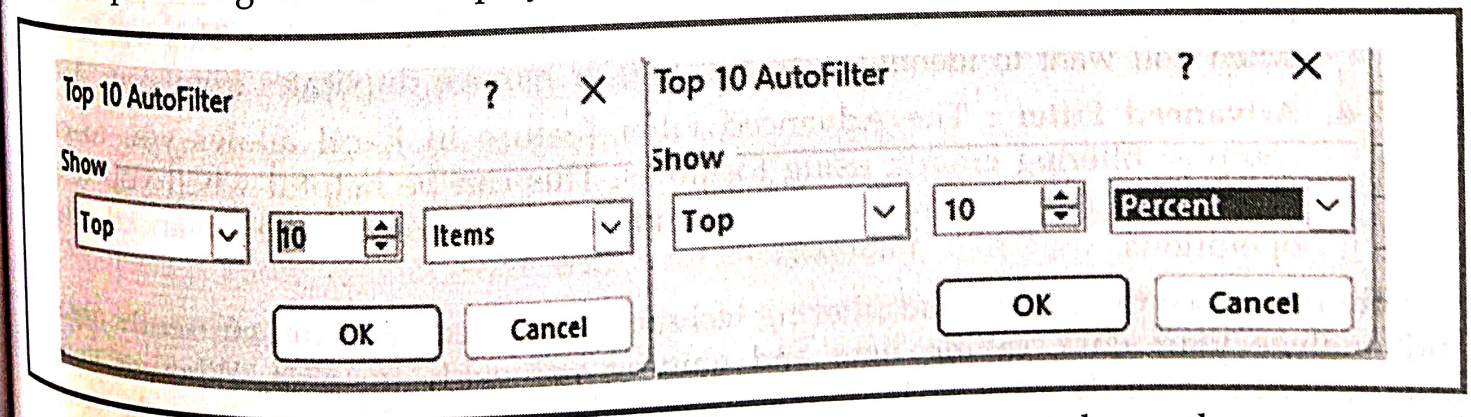
1. **Filtering by Color :** If you have applied cell formatting or conditional formatting to your table, you can filter your data based on cell color or font color. This can be useful when you want to headlight specific data points or identify patterns based on color.

2. **Filtering by Top/Bottom Values :** Excel allows you to filter your table to display the top or bottom values based on a specific number or percentage. This can be helpful when you want to identify the highest or lowest values in your dataset.

Consider the food sales data. If we want to filter the top n values or botton n values then we use this option. Select number filter and then top option from the sub menu.



A popup will appear where you can select whether top or bottom item should appear then how many items should be appear and lastly whether items or percentage should displayed.



In the following figures first figure contains top ten values where as second figure contains the first 10 percent records of the total records.

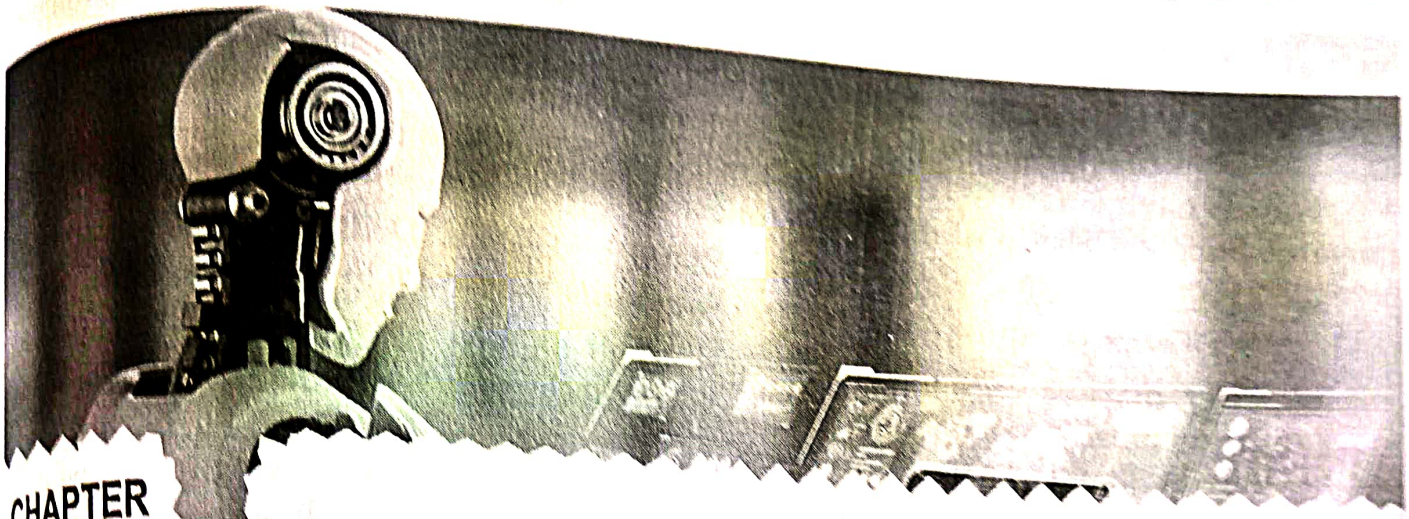
Date_k	Store_k	Produc	Transac	Dollar_	Units_s
20220921	4	20	20001385	$6.98	2
20220921	4	12	20001385	$2.49	1
20220921	4	72	20001385	$4.91	1
20220921	4	68	20001385	$3.98	1
20220921	4	3	20001385	$0.98	2
20220921	4	69	20001385	$0.98	1
20220921	4	73	20001385	$2.48	1
20220921	4	2	20001385	$4.99	1
20220921	4	1	20001385	$3.59	8
20220921	4	74	20001385	$2.99	1

1	Date_k	Store_k	Produc	Transac	Dollar_	Units_s
199	20220921	4	20	20001385	$6.98	2
200	20220921	4	12	20001385	$2.49	1
201	20220921	4	72	20001385	$4.91	1
202	20220921	4	68	20001385	$3.98	1
203	20220921	4	3	20001385	$0.98	2
204	20220921	4	69	20001385	$0.98	1
205	20220921	4	73	20001385	$2.48	1
206	20220921	4	2	20001385	$4.99	1
207	20220921	4	1	20001385	$3.59	8
208	20220921	4	74	20001385	$2.99	1
209	20220921	10	9	5282285	$24.99	1
210	20220921	10	10	5282285	$5.59	20
211	20220921	10	11	5282285	$6.99	8
212	20220921	10	12	5282285	$4.75	1
213	20220921	10	6	5282285	$7.78	1
214	20220921	10	1	5282285	$2.69	18
215	20220921	10	14	5282285	$3.99	1
216	20220921	10	15	5282285	$1.40	1
217	20220921	10	16	5282285	$1.97	1
218	20220921	10	14	5282285	$2.99	1
219	20220921	10	70	5282285	$5.99	1

3. **Filtering by Unique Values :** If you want to filter your table to display only the unique values in a column, Excel provides a built-in option for that. This will be useful when you want to identify distinct values or remove duplicates from your dataset.
4. **Advanced Filter :** The Advanced Filter feature in Excel allows you to create complex filtering criteria using formulas. This can be helpful when you need to filter your table based on multiple conditioners perform more advanced filtering operations.

By mastering the sorting and filtering techniques in Excel, you can efficiently organize and analyze your data, saving time and gaining valueable insights. Whether you are working with small datasets or large databases, these tools will help you make sense of your data and make informed decisions.





CHAPTER

7

DATA RANGE AND DATA TOOLS

CHAPTER AT A GLANCE

7.1 Introduction to Range	7.5 Data Validation
7.2 Advantages of Using Range	7.6 Removing Duplicate
7.3 Defining and Grouping Database Range	7.7 Text to Columns
7.4 Using Database Range in Formula	7.8 Flash Fill

7.1 INTRODUCTION TO RANGE

A named range is one or more cells that have been given a name. Using named ranges can make formulas easier to read and understand because one need not to manually type or select the cells in the range. Naming range is like a shortcut which can be used in different options like chart, formulas etc. Range helps to easily find the location of data in the workbook by jumping to the specific location. There are some rules to define the range.

- Space can not be given in range name.
- A cell reference can not be the name of range.
- Range name cannot start with number.
- Range name cannot start with special character.
- Range name can be have minimum 1 character and maximum 255 character long.

7.2 ADVANTAGES OF USING RANGE

Names have a variety of uses in a workbook. A name can be applied to any cell or range. Names are also useful for the following:

- Making formulas easier to understand and write. Defined names are offered in the Formula AutoComplete drop-down as you start to type a formula.
- They make it easy to quickly find a cell or select a group of cells. That makes them ideal for navigating large worksheets or for applying formatting to cell ranges that frequently change.
- Forcing a formula reference to remain absolute, without having to use the dollar sign.
- Doing a two-way lookup with the intersection operator
- Improving report results from Solver and Scenario Manager.
- Storing a value that will be used repeatedly but that might occasionally need to change, such as a sales tax rate.
- They add an extra layer between your formulas and your worksheet. If you change the structure of your worksheet, you don't need to modify your formulas. Instead, you simply edit the named ranges so they point to the new cell locations.

7.3 DEFINING AND GROUPING DATABASE RANGE

Excel offers the different ways to define the name range. Following section discuss the common ways to define range using different option.

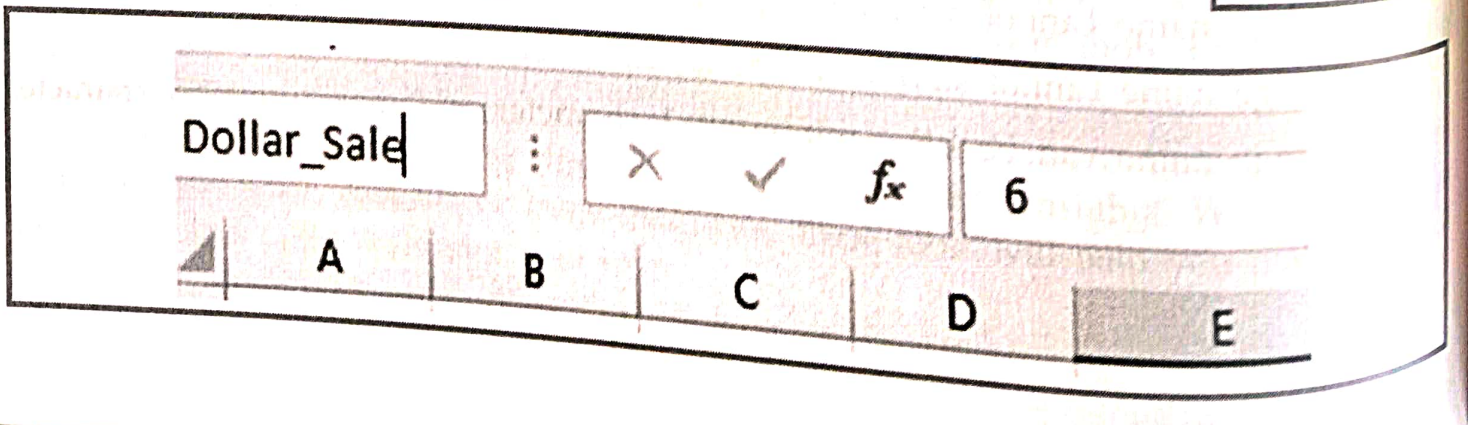
Dollar_sales
$6.00
$3.00
$3.49
$2.49
$2.69
$11.99
$6.00
$2.48
$9.81

7.3.1 DEFINING NAME USING NAME BOX

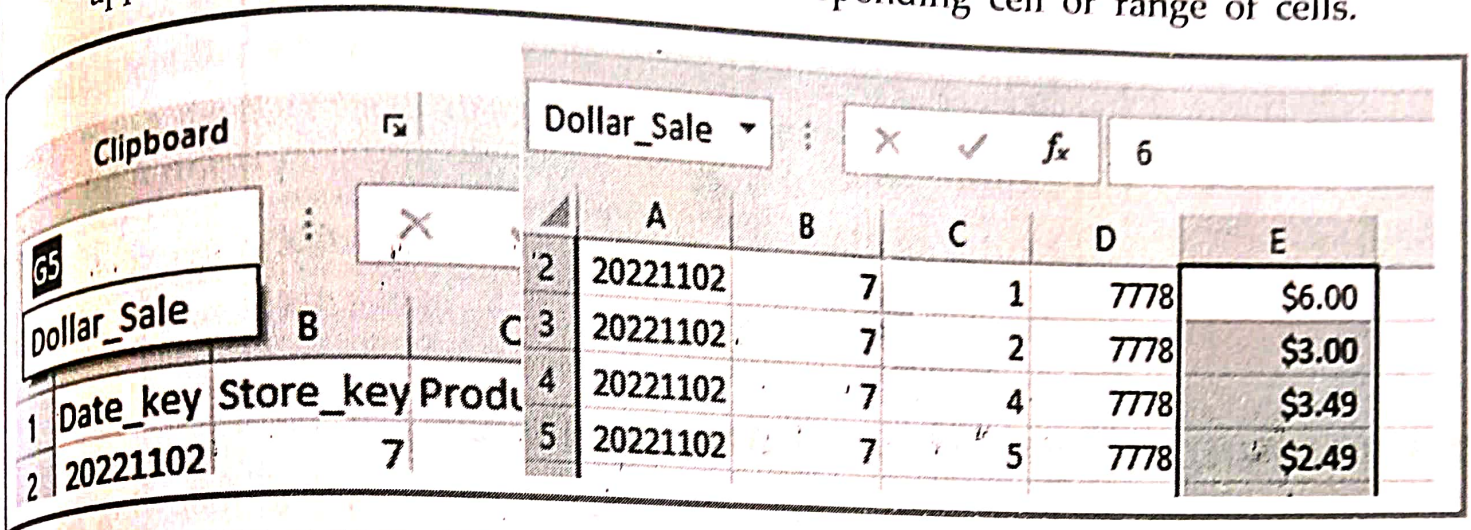
Left end of the formula bar has rectangle which is called as name box it displays the address of the selected cell generally. We can define the range name using this name box. Following steps discuss the how to define the name of the range using name box.

1. Select the column values range for which name has to be defined. In this example Dollar_Sales column is selected having range from E2:E650

2. Type the range name as Dollar_Sale in the name box as show below and press enter.



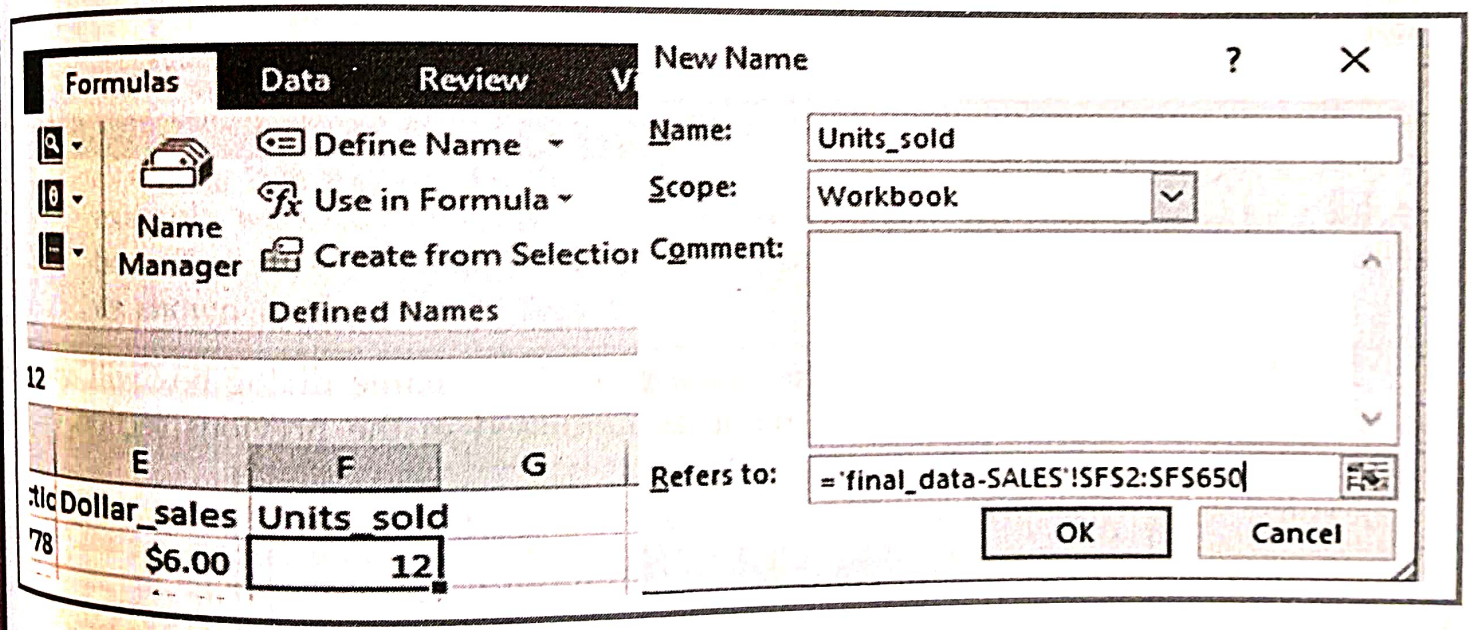
3. You can use this defined name any time using this name box. To use a name, click the drop-down arrow to the right side of the name box to see the list of names defined in the workbook. When you click a name, Excel jumps to the appropriate position and selects the corresponding cell or range of cells.



7.3.2 DEFINING NAME USING FORMULA BAR

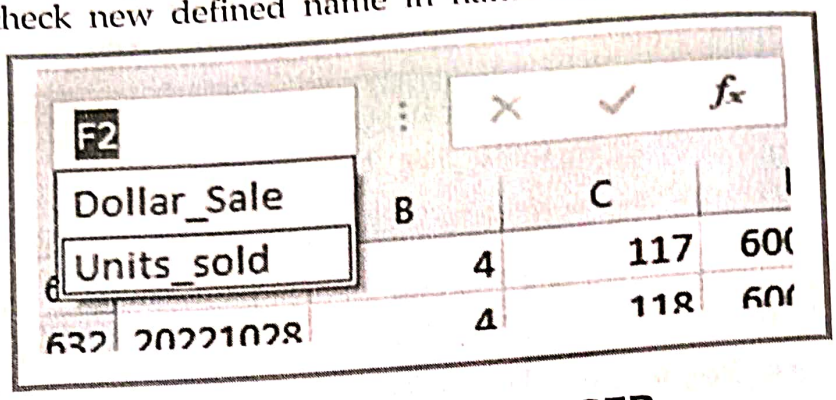
The Formulas tab contains a group called Defined Names. The following example introduces the Name dialog.

1. Click on the cell you would like to name. Then click on the define name option in formula menu as show below. Define name dialog box will appear.



2. Define name box has the following fields:
   - **Name** is the name you give the range (like Units_sold).
   - **Scope** lets you control where you can use the name. Ordinarily, names have Workbook scope.
   - **Comment** is a description for your named range. This description appears in a tooltip when you enter the name into a formula.
   - **Refers** to is a cell reference that indicates the cells you selected.

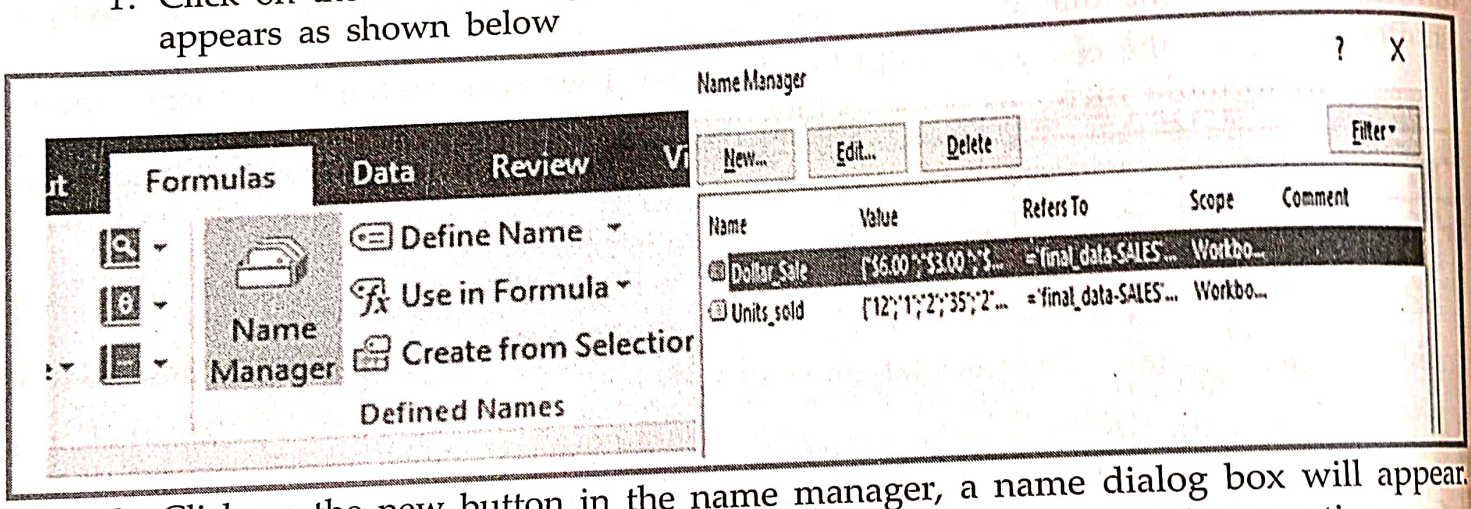
3. Now select the range for which you want define name using Refer to column as shown in the figure in step 1. Press ok and name is defined in the dialog box. You can check new defined name in name box drop down as shown below



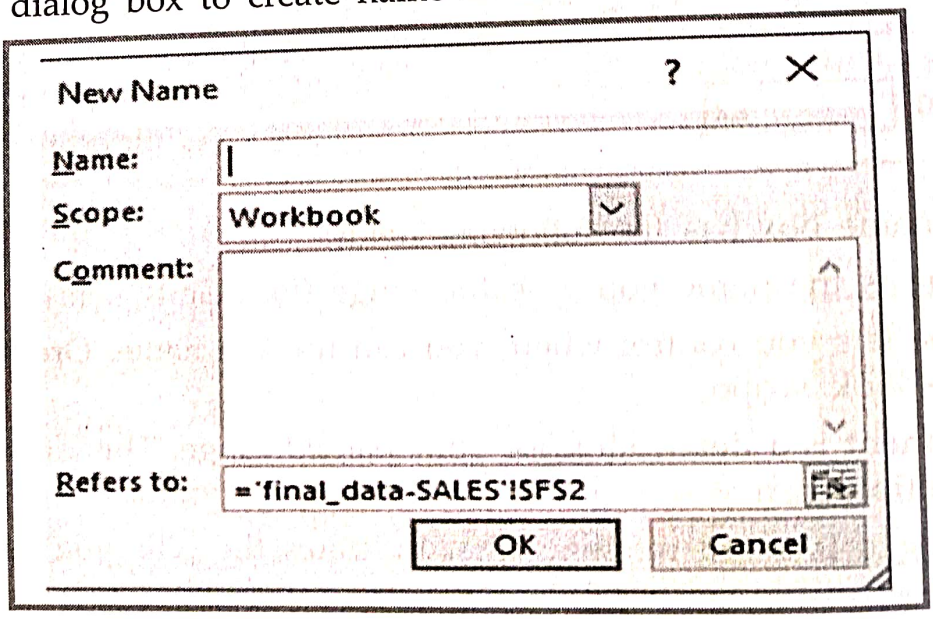
7.3.3 DEFINING NAME USING NAME MANAGER

The Formulas tab contains a group called Name Manager. The following example introduces the Name Manager

1. Click on the name manager in the formula tab then a name manager window appears as shown below



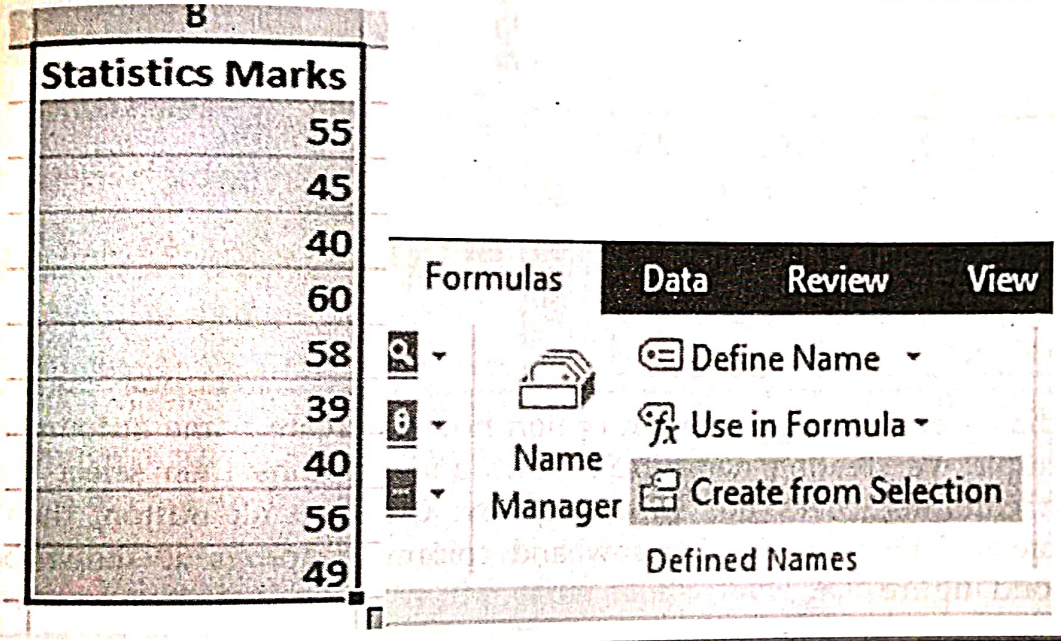
2. Click on the new button in the name manager, a name dialog box will appear. Use this dialog box to create name as discussed in the previous section.



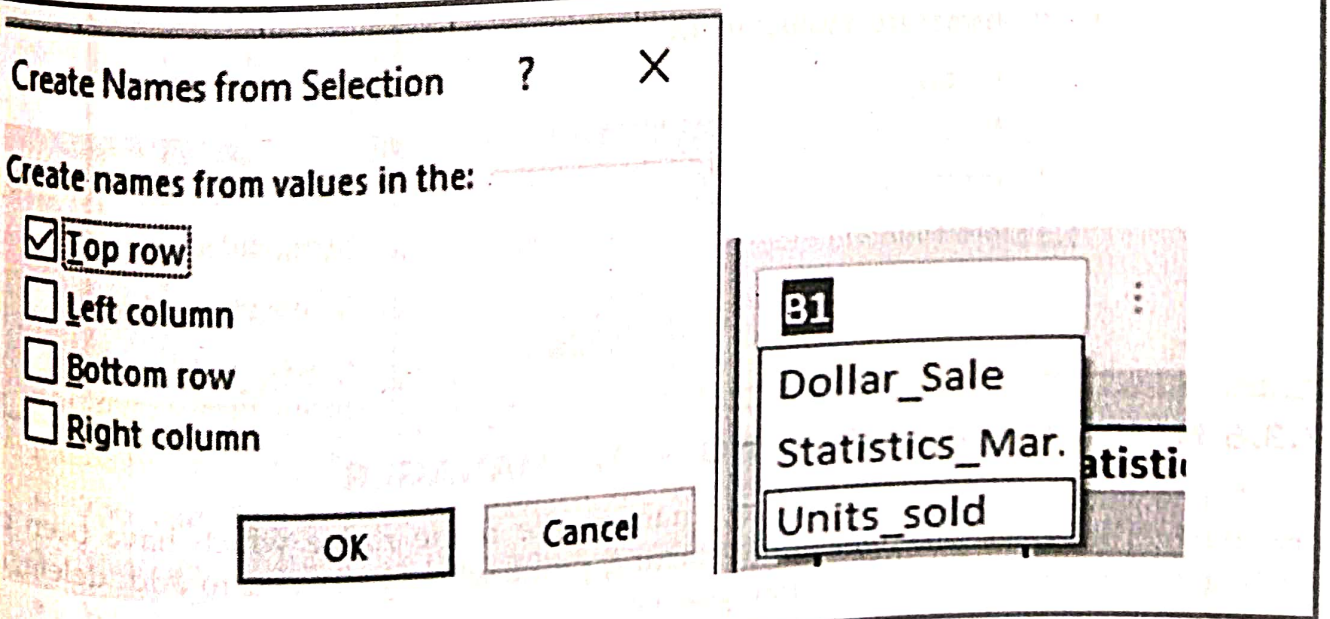
1.3.4 DEFINING NAME RANGE FROM SELECTION (SINGLE ROW OR SINGLE COLUMN)

Excel also has the built-in smarts to automatically generate names for cell ranges. To perform this trick, Excel searches a group of cells you select and, with a little help from you, identifies which cell or cells would offer appropriate names.

1. Select the data for which you would like to create name range. Click on 'create from selection' in the formula tab as show below



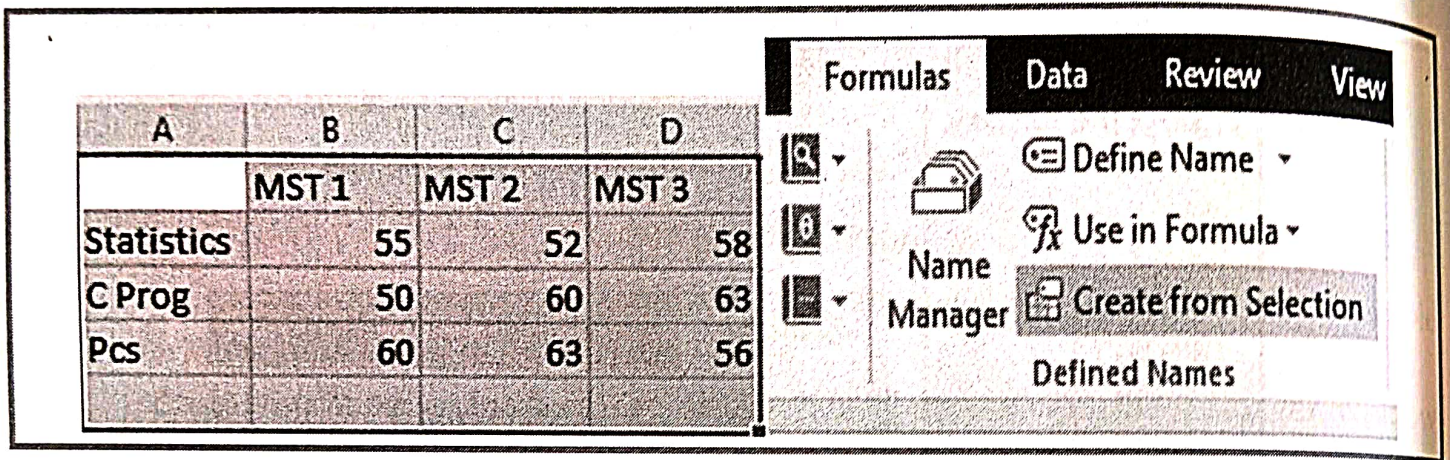
2. A dialog box will appear with option how to create name automatically. In this example we creating name from row then select 'top row' check box in dialog box. Click on ok button, then new name range will be created using row heading as shown in the second figure.



7.3.5 DEFINING NAME RANGE FROM SELECTION (ROWS AND COLUMNS OR TWO-WAY TABLE)

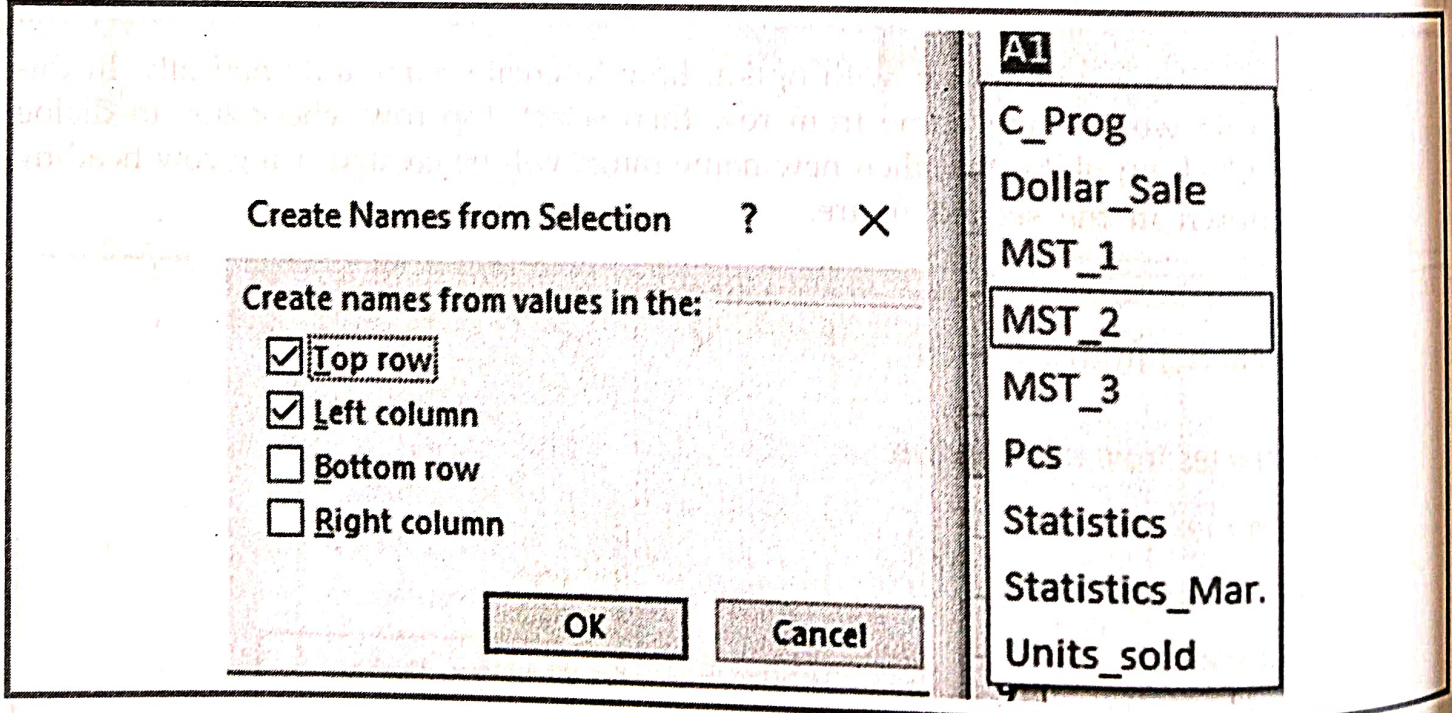
In Excel, you can apply names by using both the row labels and column headings at the same time. In Figure, the selections in the Create Names from Selected Ranges dialog mean that six new names will be added to the workbook.

1. Select the data for which you would like to create name range. Click on 'create from selection' in the formula tab as show below



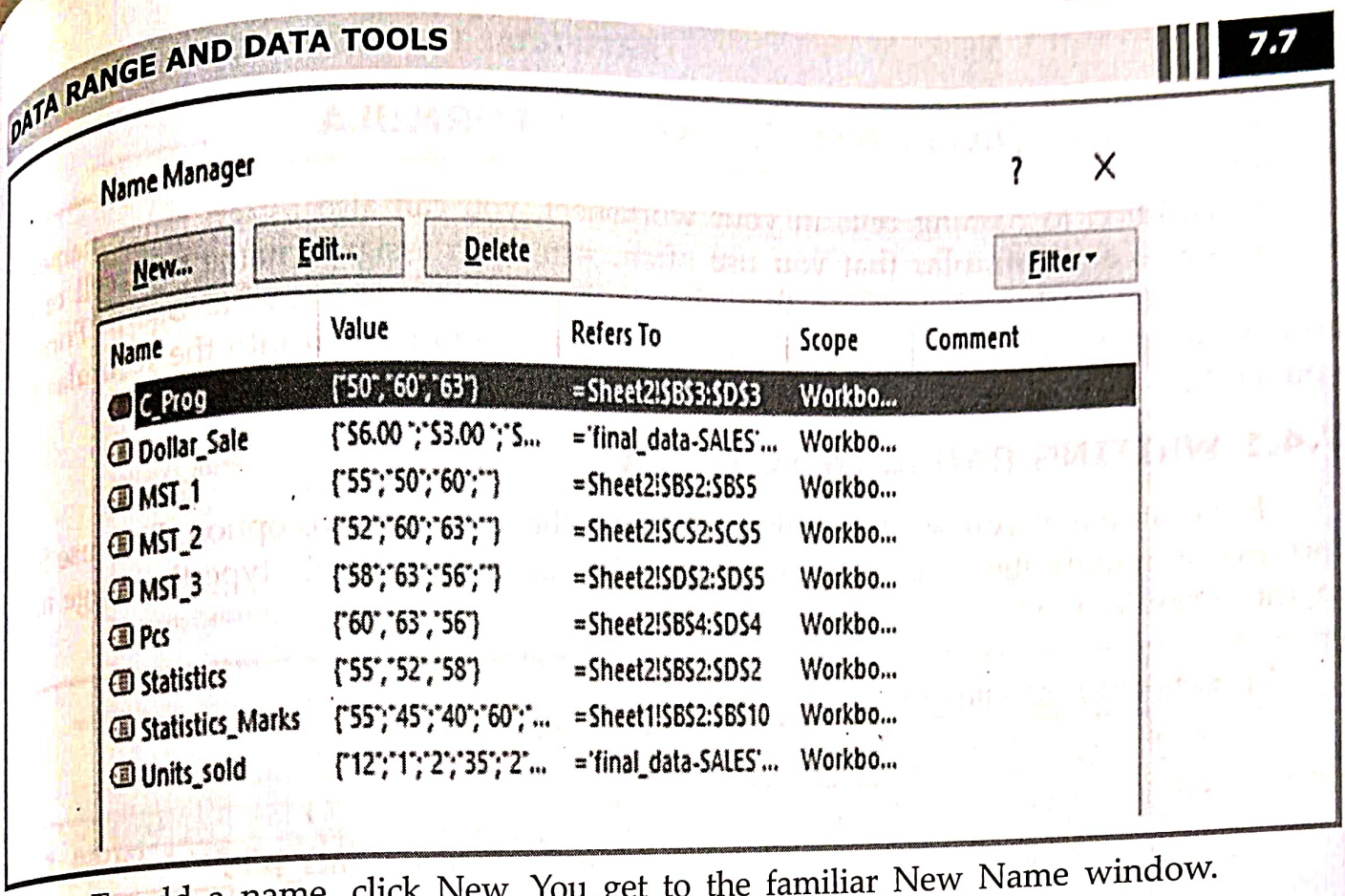
	A	B	C	D
		MST 1	MST 2	MST 3
	Statistics	55	52	58
	C Prog	50	60	63
	Pcs	60	63	56

2. A dialog box will appear with option how to create name automatically. In this example we creating name from rows and columns then select 'top row' and 'left columns' check box in dialog box. Click on ok button, then new name range will be created using row and column headings as shown below in the second figure.

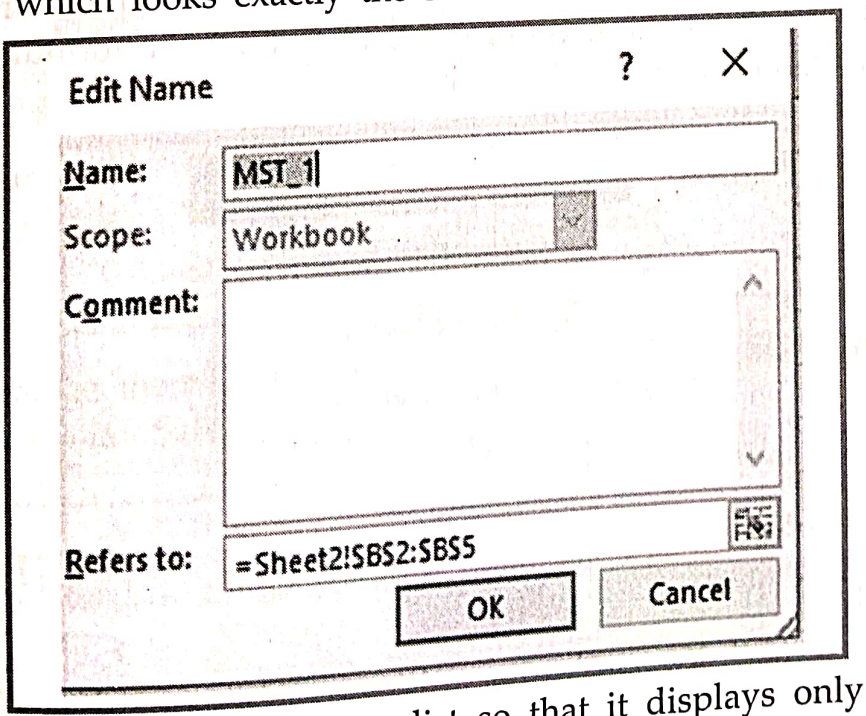


7.3.6 MANAGING NAME USING NAME MANAGER

Name manger in the formula tab manage the name range which have been created previously. The Name Manager is the starting point when you need to add, delete, or edit existing named ranges. Here's what you can do:



- To add a name, click New. You get to the familiar New Name window.
- To remove a name, select it in the list, and then click Delete. It will remove the range reference in the formula as well if used any.
- To edit a name, select it in the list, and then click Edit. You see an Edit Name window, which looks exactly the same as the New Name window.



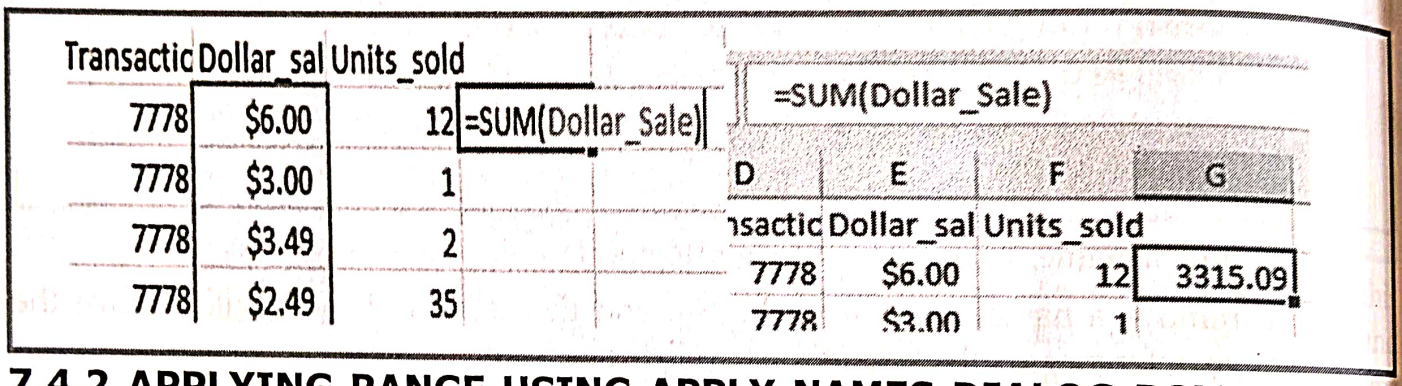
- Filtering cuts down the names list so that it displays only the names you're interested in. The Name Manager lets you use several types of filters:
  - You can find names limited to the current worksheet.
  - You can find names that point to cells with errors (and those that don't).
  - You can find ordinary names, or names used in tables

7.4 USING DATABASE RANGE IN FORMULA

In addition to naming cells in your worksheet, you can also assign range names to the constants and formulas that you use often. After you assign a name to a cell or cell range in your worksheet, you can then choose the range name from the Use in Formula button's drop-down menu on the Ribbon's Formulas tab to paste it into the formulas that you build.

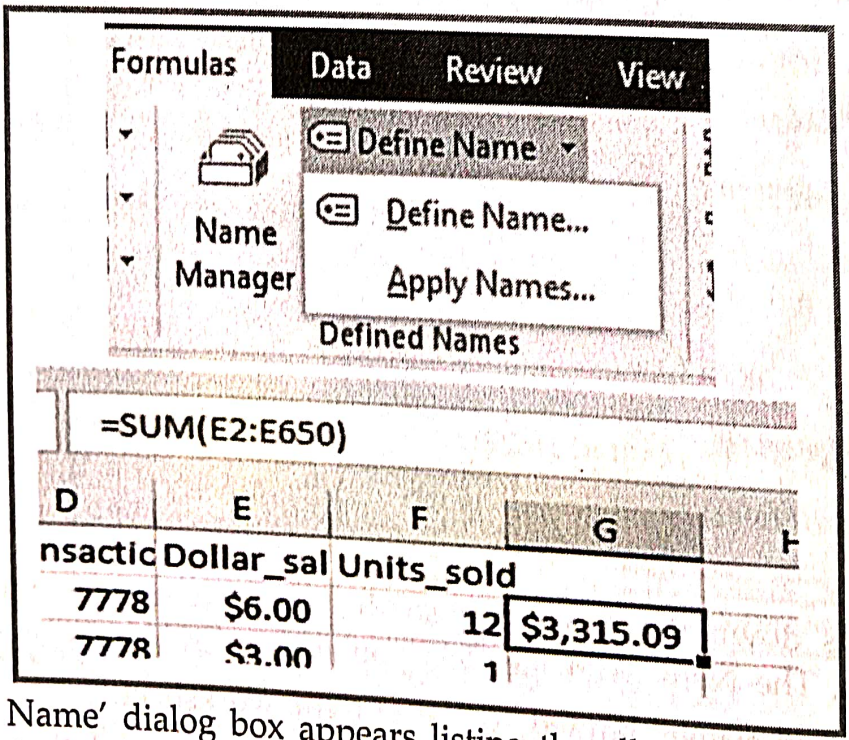
7.4.1 WRITING RANGE IN FORMULA

For example, if you want to sum all the sales the you have two option. First use SUM formula by adding the range manually like SUM(E2:E650) or simply type the range name in the formula as shown below

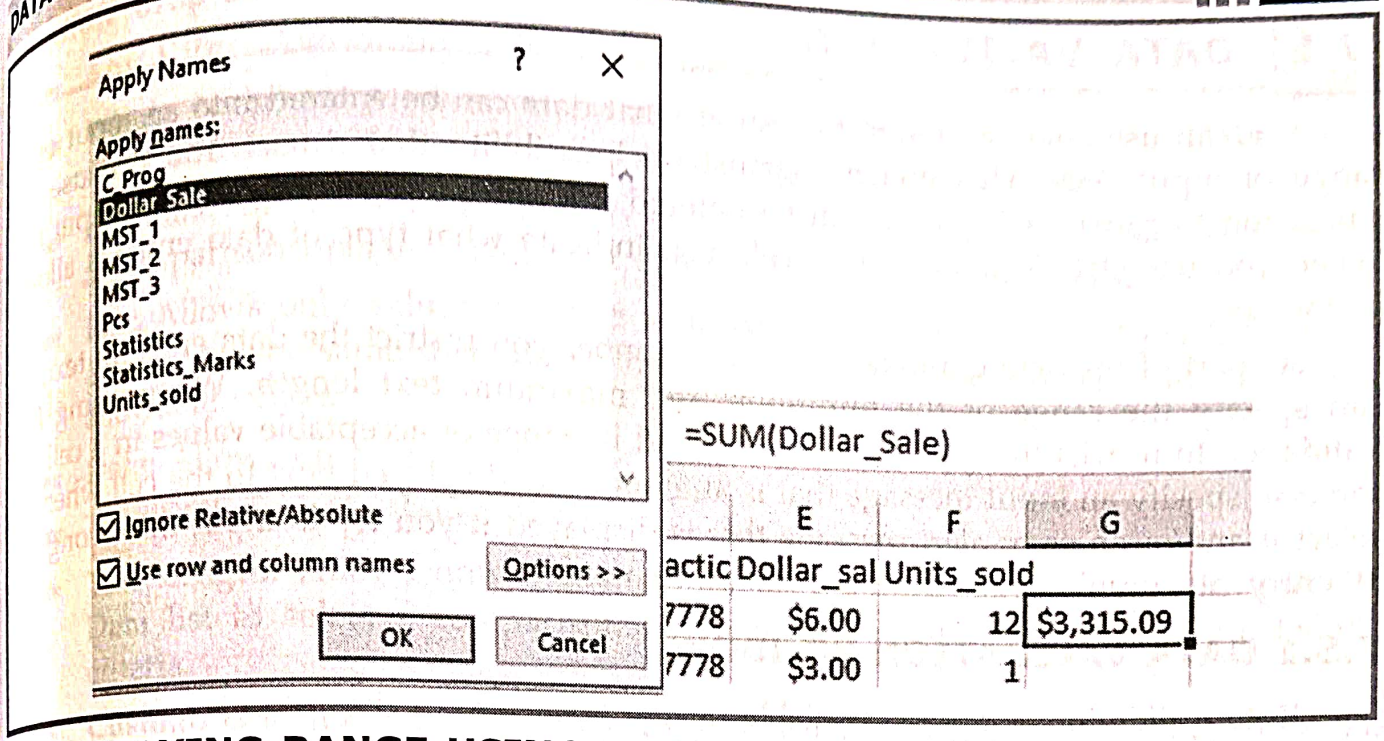


7.4.2 APPLYING RANGE USING APPLY NAMES DIALOG BOX

If you have defined many name ranges, then it is difficult to remember all the name ranges and apply them in formula. Excel provides two options in formula tab. First option 'apply name' in the define name as shown below

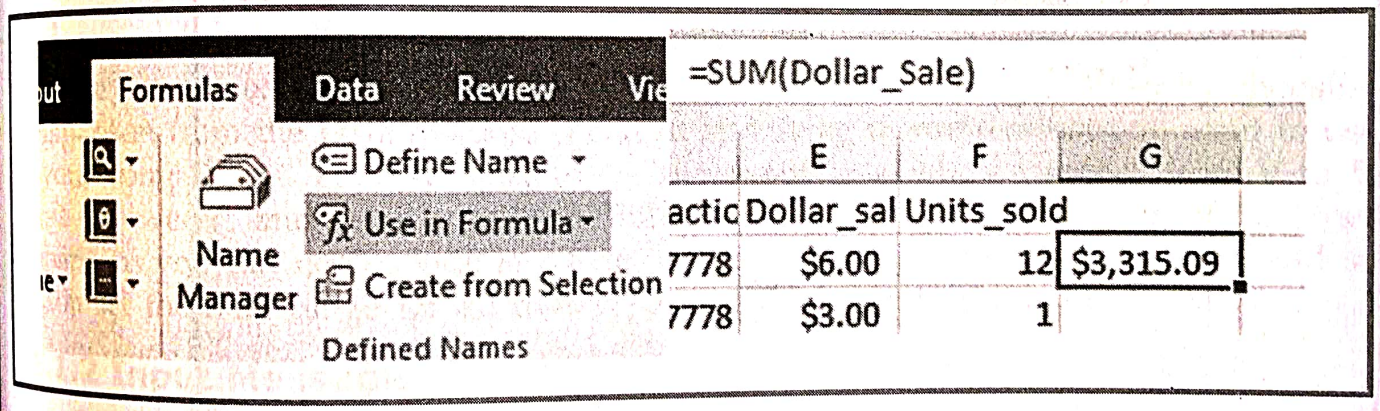


An 'Apply Name' dialog box appears listing the all the available name range. Select name range and press ok. Excel replace the given range is SUM function from SUM (E2:E650) to SUM(Dollar_Sale).

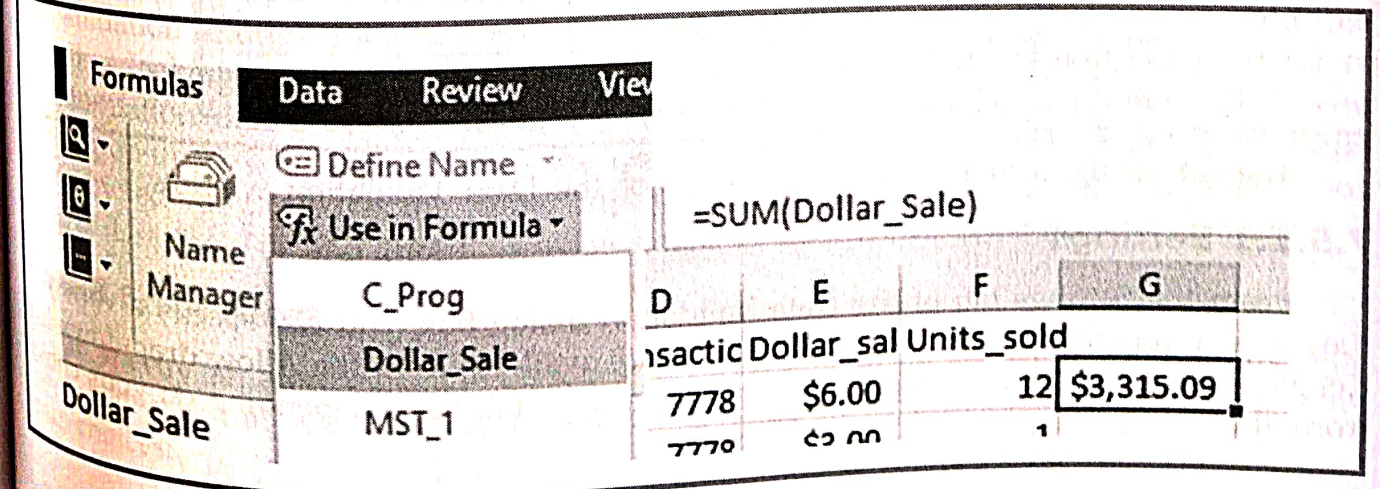


7.4.3 APPLYING RANGE USING USE IN FORMULAS OPTIONS

Second option is 'Use in Formula' which listing the already defined range as a drop down box.



Click on the range then Excel replace the given range in SUM function from SUM (E2:E650) to SUM(Dollar_Sale).



7.5 DATA VALIDATION

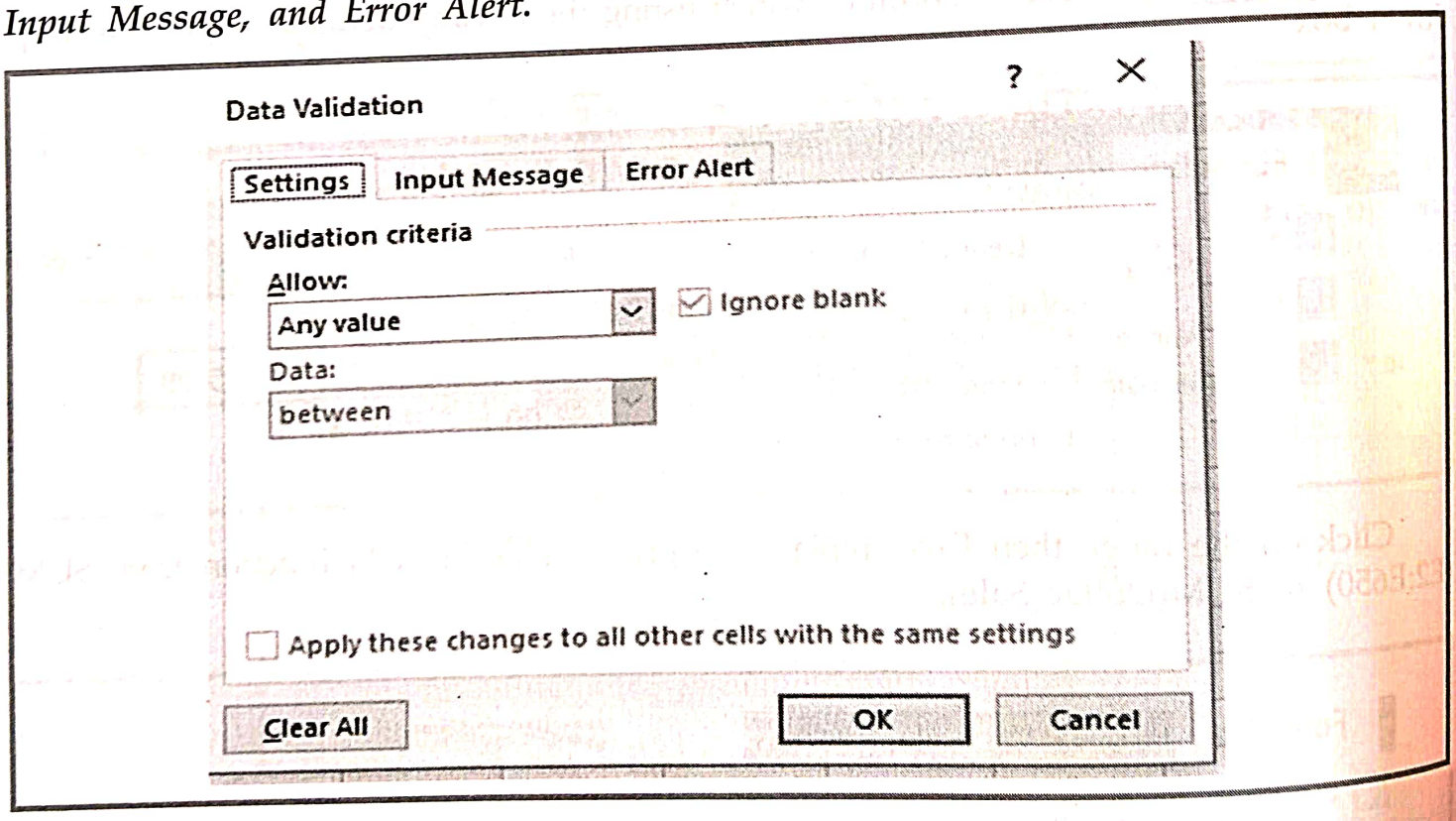
You can use data validation to restrict what data can be entered into an input cell or range of input cells. This guards against the possibility that a user You can use data validation to restrict what data can be entered into an input cell or range of input cells. When you use Data Validation in a cell, you indicate what type of data entry is allowed in the cell.

As part of restricting a data entry to a number you restrict the data entry to text, you can specify the range of the minimum and maximum text length. When using Data Validation to restrict the type of data entry and its range of acceptable values in a cell, you can also specify an input message that is automatically displayed next to the cell when you select it and/or, an error alert message that is displayed if you try to input the wrong type of entry ora number outside the permissible range.

7.5.1 DATA VALIDATION WINDOW

Data Validation option is available on the Data tab, in the

Data Tools group. In the Data Validation dialog box, there are three tabs: Settings, *Input Message, and Error Alert.*



The following sections explain each tab of the Data Validation window.

7.5.1.1 Settings

From the Settings tab of the Data Validation dialog box, you can specify a variety of data validation criteria. The following options are available from the Allow drop-down list. Keeping mind that the other controls on the Settings tab vary, depending on your choice from the

Allow drop-down list.

- **Any Value:** Allows any value in the range.
- **Whole Number:** Allows only whole numbers (integers). Use the Data list to select a comparison operator and then enter the specific criteria.
- **Decimal:** Allows decimal numbers or whole numbers. Use the Data list to select a comparison operator, and then enter the specific numeric criteria.
- **List:** Allows only values specified in a list. Use the Source box to specify either a range on the same sheet or a range name on any sheet that contains the list of allowable values.
- **Date:** Allows only dates.
- **Time:** Allows only times.
- **Text Length:** Allows only alphanumeric strings of a specified length. Use the Data list to select a comparison operator, and then enter the specific length criteria.
- **Custom:** Use this option to enter a formula that specifies the validation criteria. You can either enter the formula directly into the Formula box etc.

To allow blank entries, either in the cell itself or in other cells specified as part of the validation settings, leave the Ignore Blank check box selected.

Data

Once you choose the data type, you need to set the data range. Do so by choosing a comparison from the Data list box, and then specifying the values that you want to use for your comparison. For example, to restrict input to a whole number from 0 to 5, choose "between" and set a minimum value of 0 and a maximum value of 5. Other comparisons you can use include less than, greater than, greater than or equal to, and so on. You set the data range for all data types, including dates, times, and text

7.5.1.2 Input Message

An input message has two components: a title (displayed in bold) and a more detailed message. You enter both of these pieces of information in the Input Message tab of the Data Validation window. Input messages should contain more than a description of your data validation settings. Ideally, your input message explains a little bit about the data that the cell is looking for. For example, you may decide to describe the type and format of the information that the cell accepts, and the restrictions you placed on the cell. If you want a pop-up box to appear when the user selects the restricted cell or any cell within the restricted range, leave the Show Input Message When Cell Is Selected check box selected. Use the Till and Input Message boxes to specify the message that appears.

7.5.1.3 Error Alert

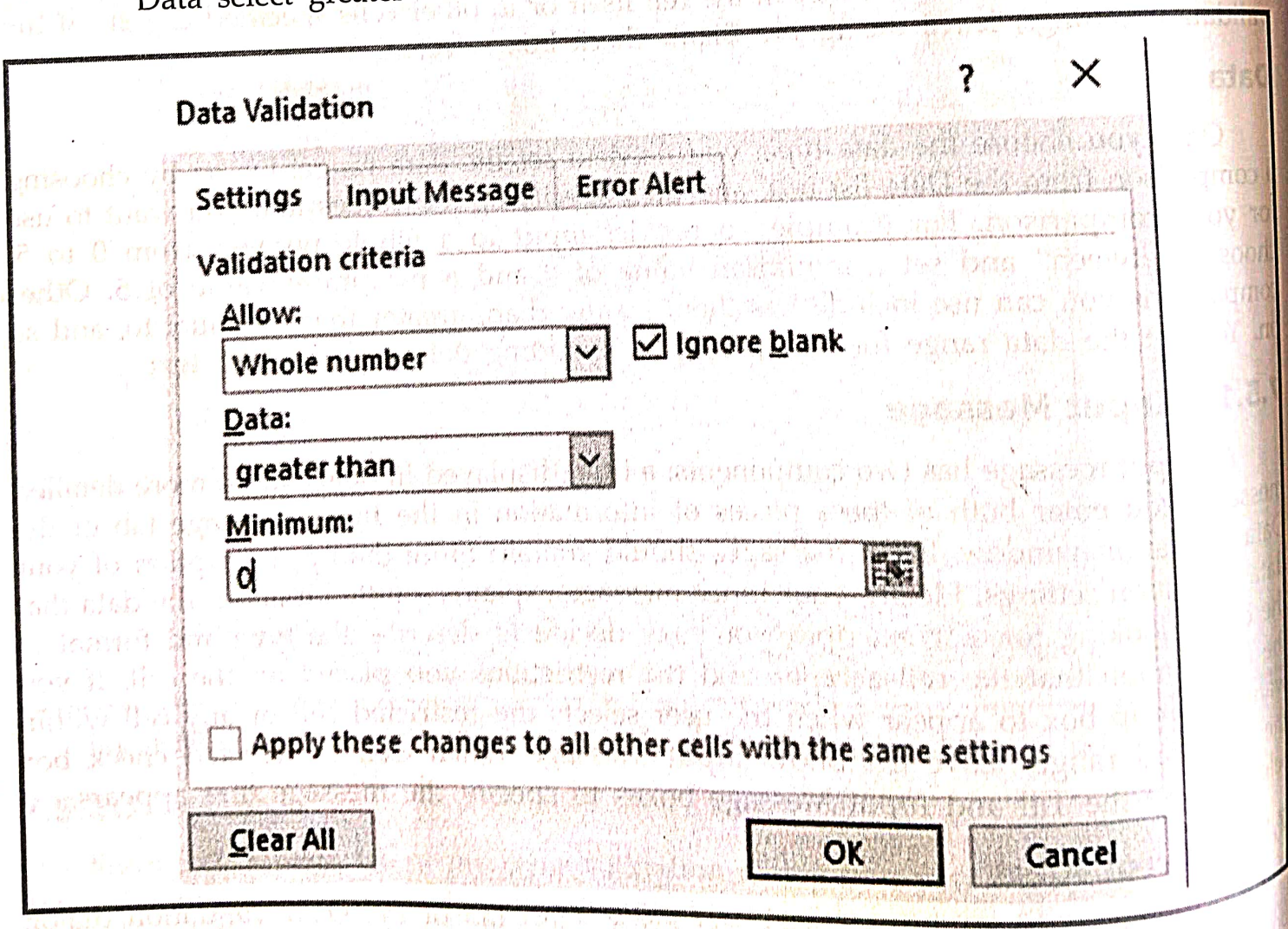
To specify an alert message, click the Error Alert tab of the Data Validation dialog box, where you can select the kind of warning from the Style drop down list. Here are your choices:

- **Stop:** Excel displays the error message along with a Retry and Cancel button. The person using the workbook must click Cancel to reverse the change. he Stop option is the only style choice that completely prevents the person using the workbook from entering invalid data.
- **Warning:** In this case, the error message includes Yes and No buttons that let the person entering the data decide whether to go ahead with their input. Clicking Yes makes Excel accept the data entered into the cell, even if it breaks the validation rules.
- **Information:** This error message comes with Cancel and OK buttons. Clicking OK enters the new (invalid) data in the cell; Cancel leaves the cell unchanged. After you set the icon, specify a title and then type a descriptive error message.

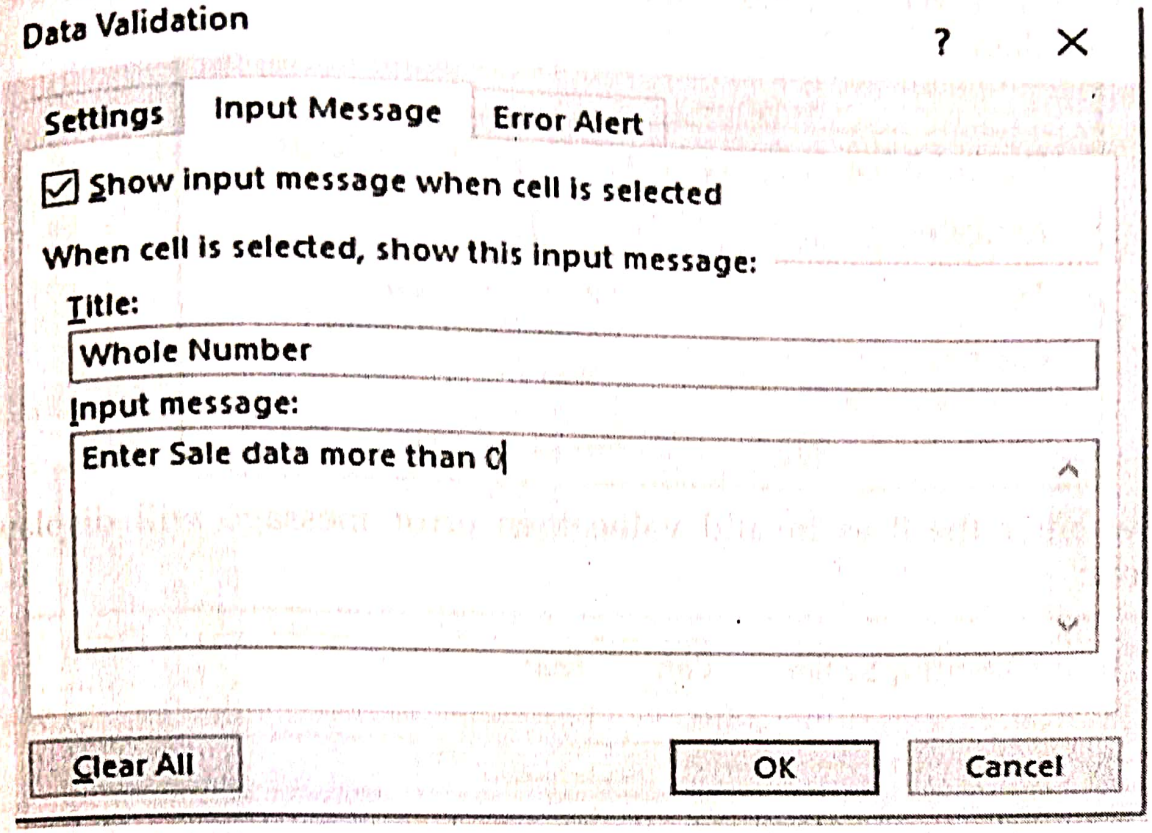
7.5.2 VALIDATE WHOLE NUMBER

We want to enter the sales data for multiple products and want to restrict users from entering zero or negative values. Select the cells you want to add data validation to.

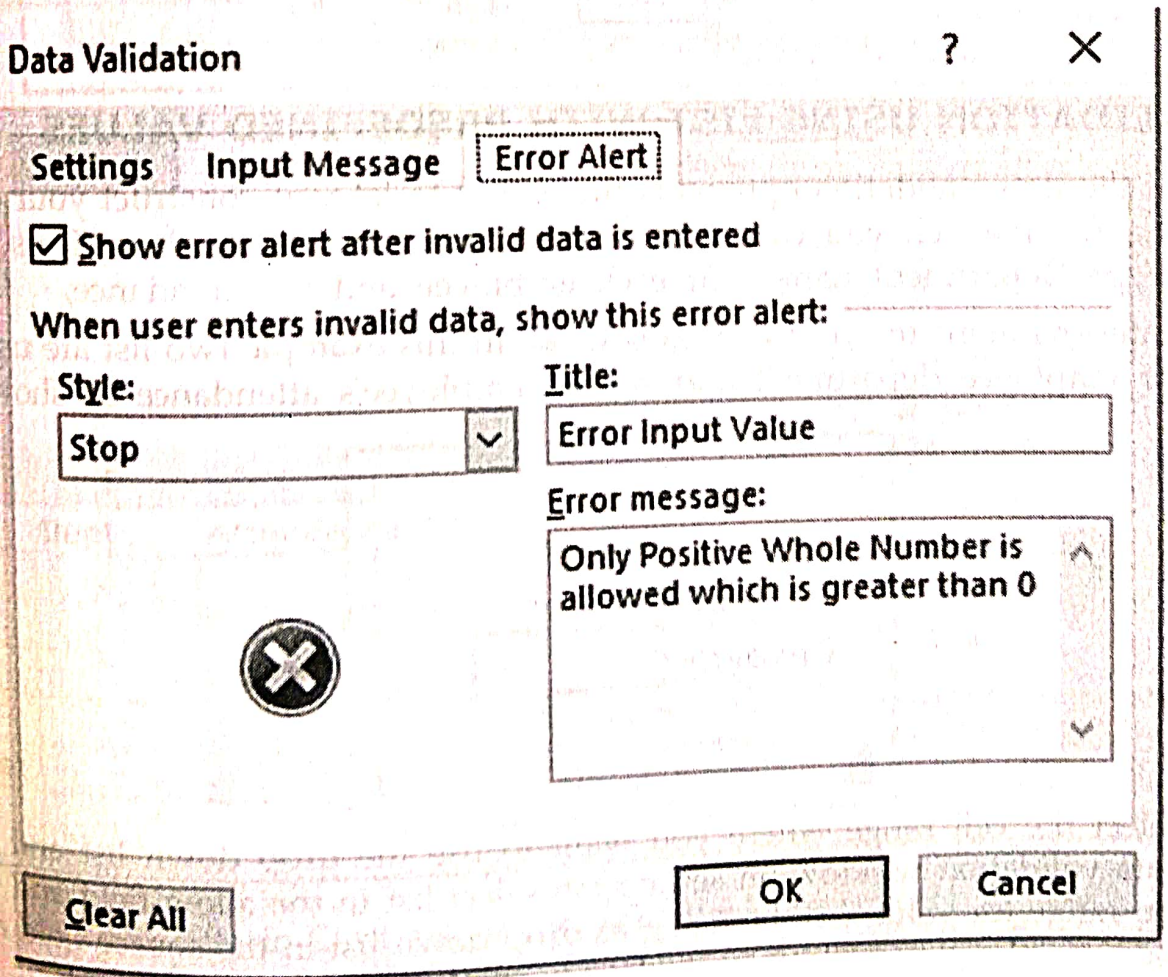
1. Select cell for which validation is required and Open the Data Validation dialog box. In the Settings tab, under Allow, select the Whole Numbers option and in Data select greater than with Minimum value as 0.



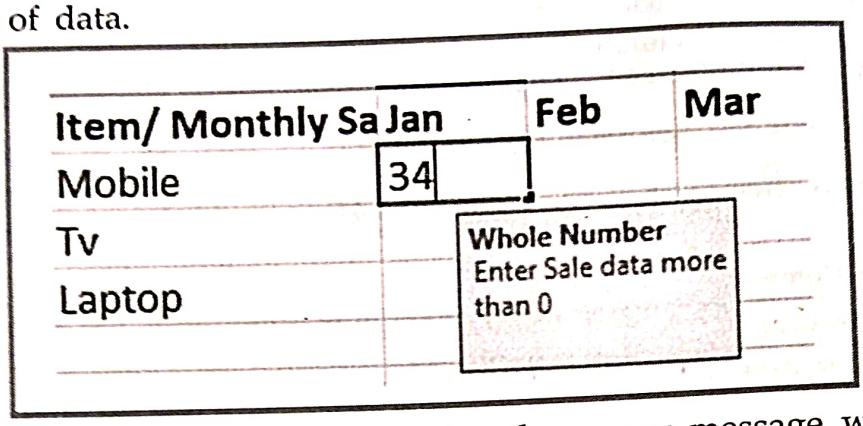
2. Set the input message tab to specify hint for input data.



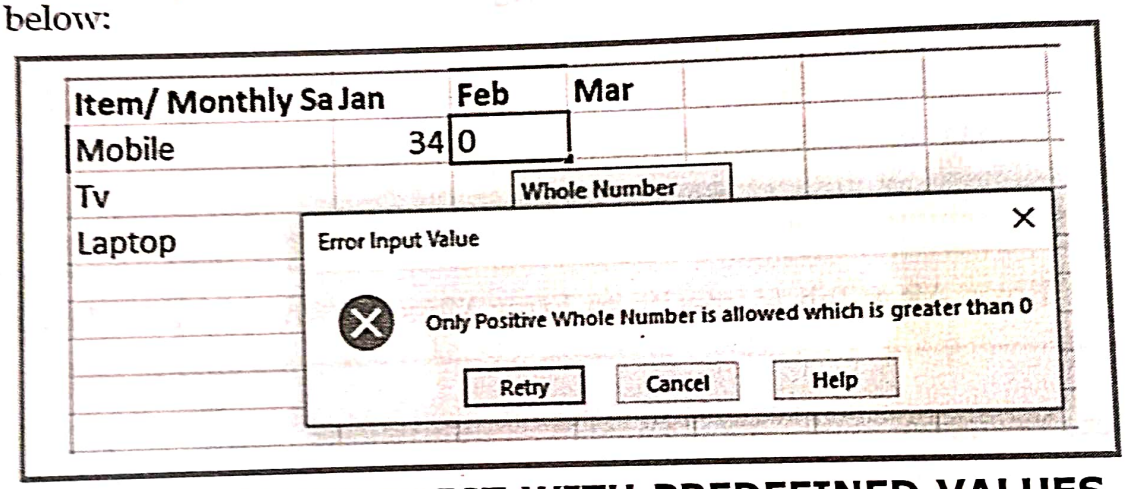
3. Select Error Alter Tab to specify to Error Message when invalid data is entered.



4. Now try to enter the valid value for the sale, input message hints the user which type of data.



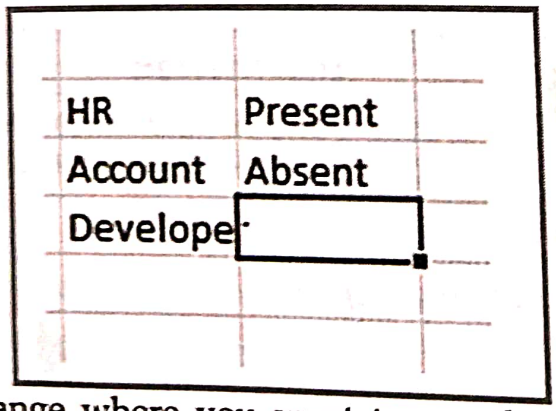
5. Now enter the 0 as invalid value then error message will display as shown below:



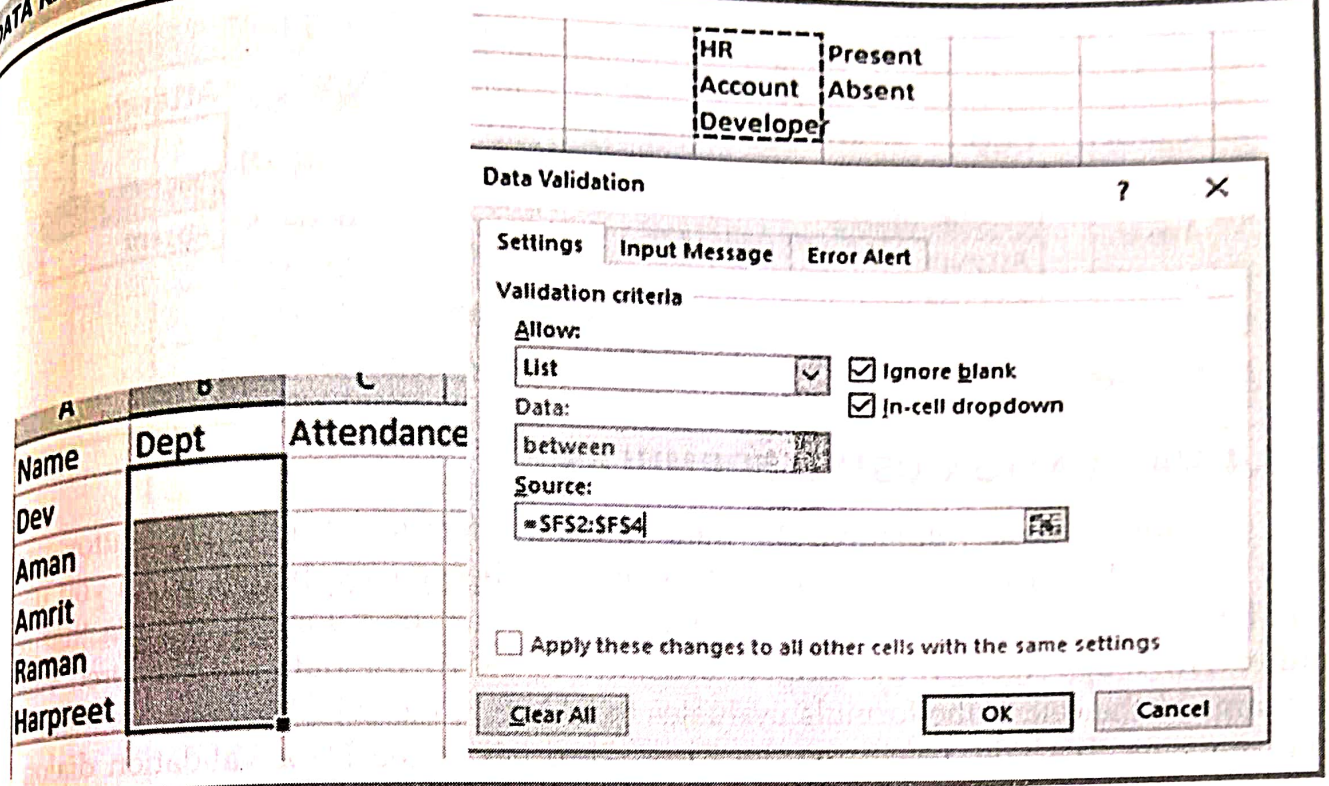
7.5.3 VALIDATION USING LIST WITH PREDEFINED VALUES

One of the data validation options is the ability to let you construct your own drop-down menu, from which you can choose from your own list of data. Thus instead of having to type Department names for each employee and its attendance,

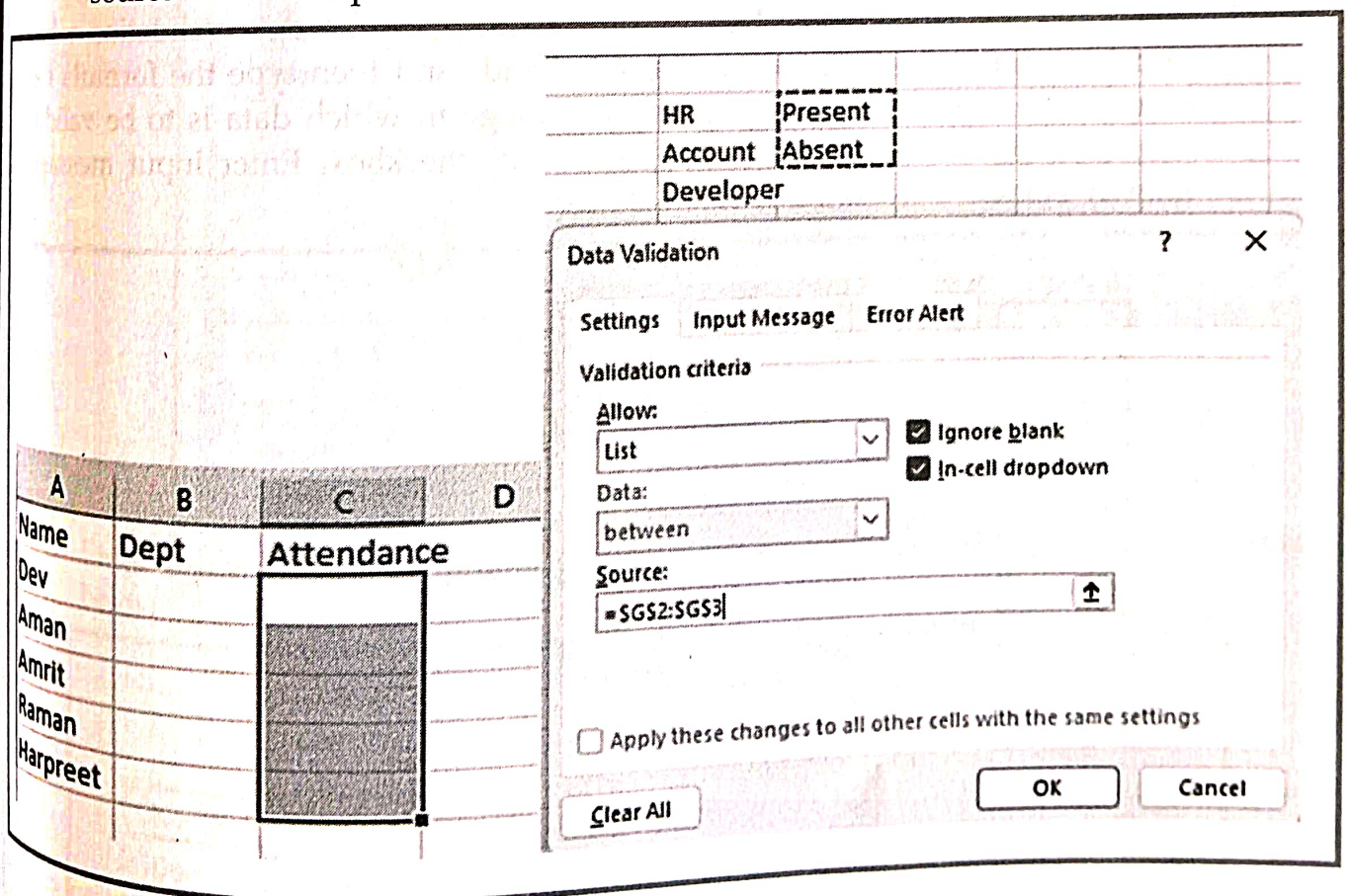
1. Enter the items for your drop-down list. In this example two list are used i.e. one for employee department and second employee's attendance as shown below:



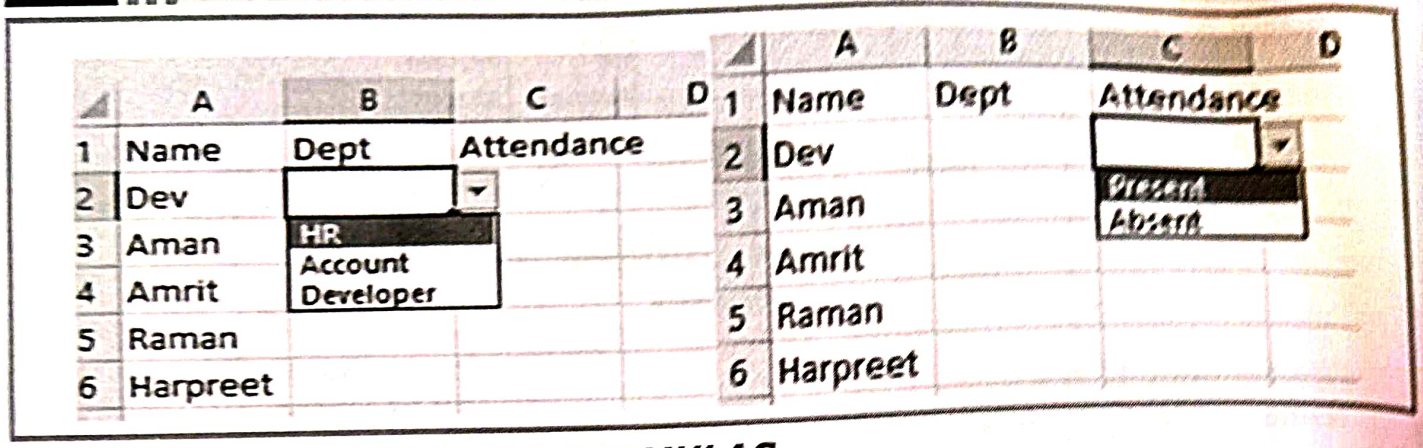
2. Select the cell range where you want to use the drop down list and the open data validation window. In setting tab select list in the allow field and the select the range for value will be using as drop down list in the source field and press ok.



3. Next select Attendance column cells which will be used for data entry and open data validation window and select list in allow field and range of list values in source field and press ok.



4. Enter the data in Dept and Attendance field respectively. For both columns two lists will be displayed to enter the value in the selected cell.

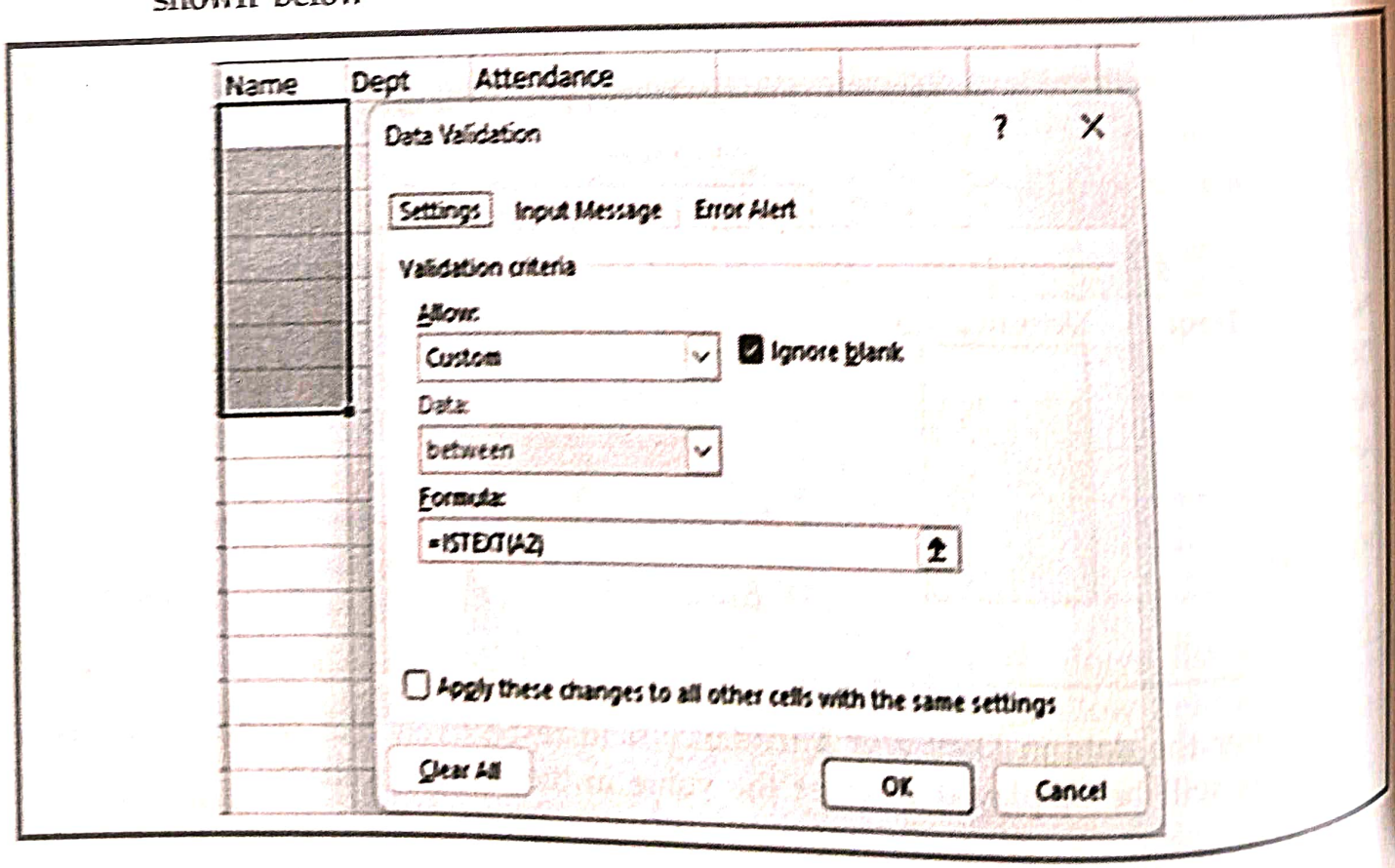


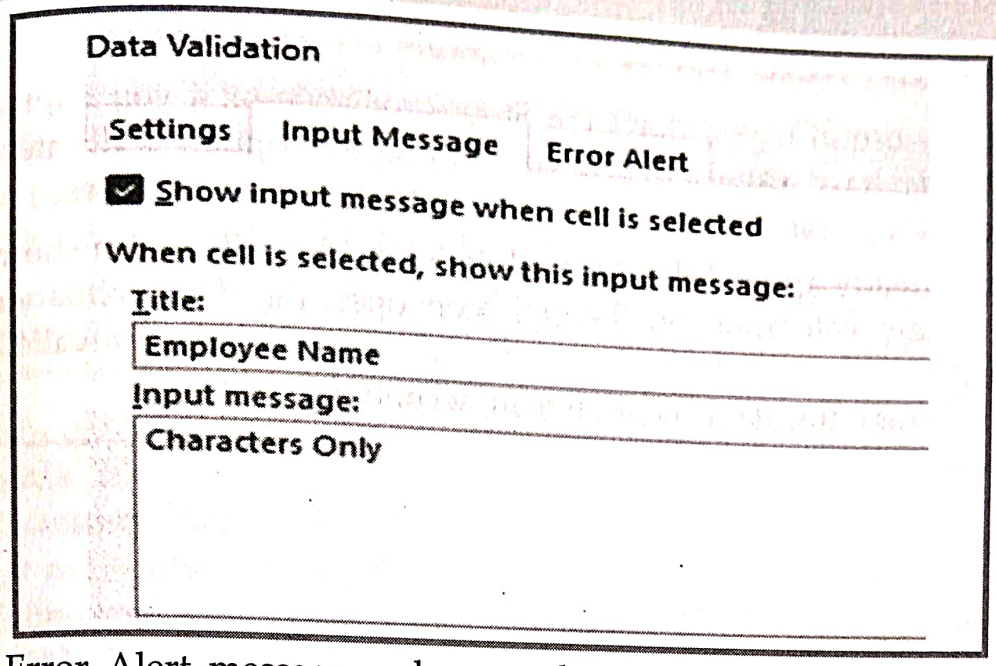
7.5.4 VALIDATION USING FORMULAS

For simple data validation, the data validation feature is quite straightforward and easy to use. The real power of this feature, though, becomes apparent when you use data validation formulas. The formula that you specify must be a logical formula that returns either TRUE or FALSE. If the formula evaluates to TRUE, the data is considered valid and remains in the cell. If the formula evaluates to FALSE, a message box appears that displays the message that you specify on the Error Alert tab of the Data Validation dialog box.

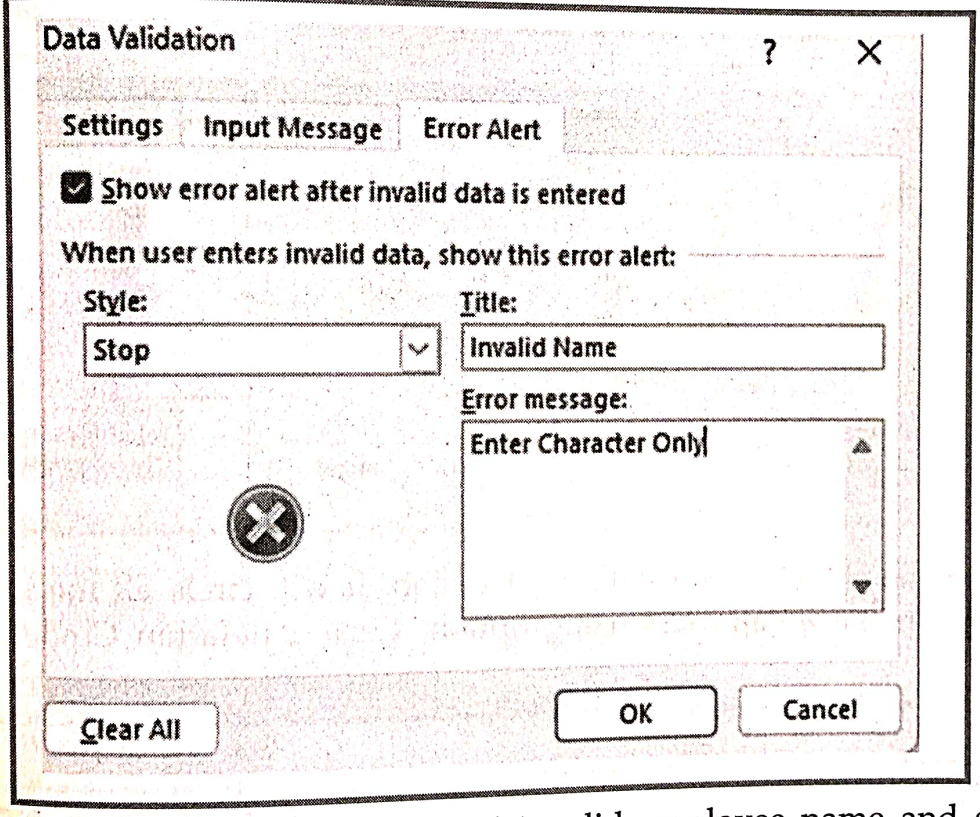
It is mandatory that employee name must be in characters then we can apply this type of data validation using formula in the data validation window. We will use previous data and assume that validation is done on name column

1. Select the cell range of the Name column and open the data validation window and in setting tab select custom in allow fields and then type the formula in the formula field. In the formula specify cell range in which data is to be validated. Select 'Apply these changes to all other cell' checkbox. Enter input message as shown below

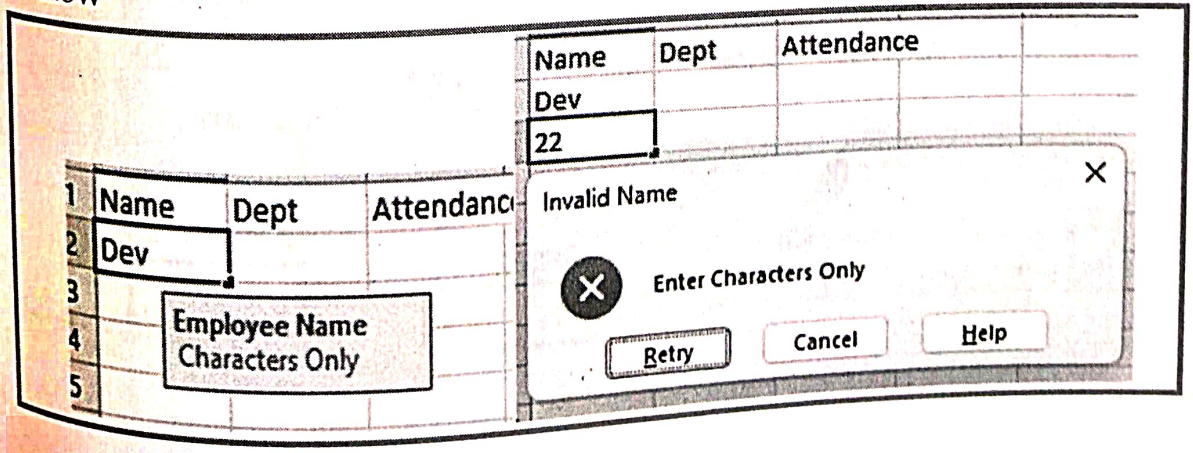




2. Enter Error Alert message and press ok.



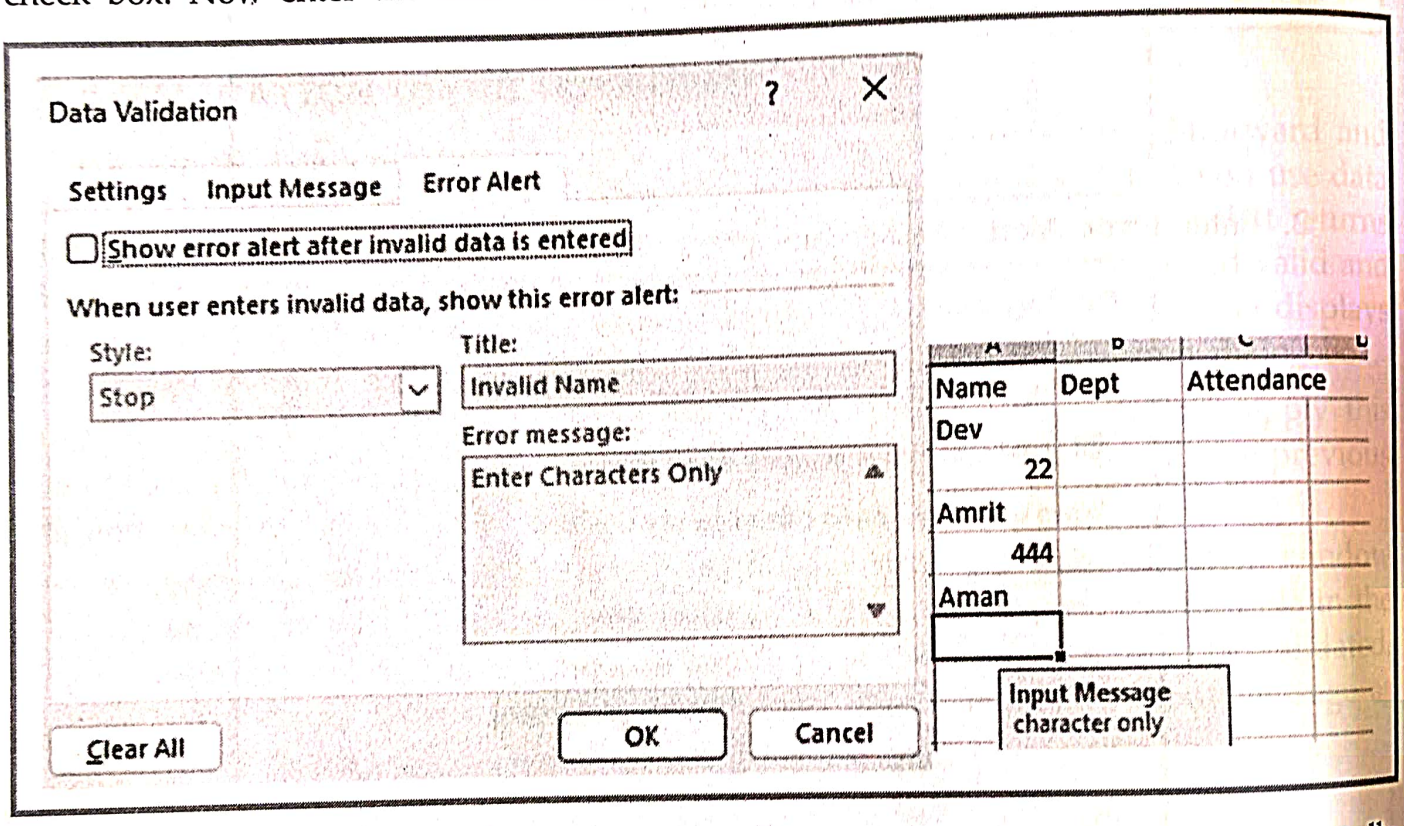
3. Enter the valid employee name and invalid employee name and see the result below



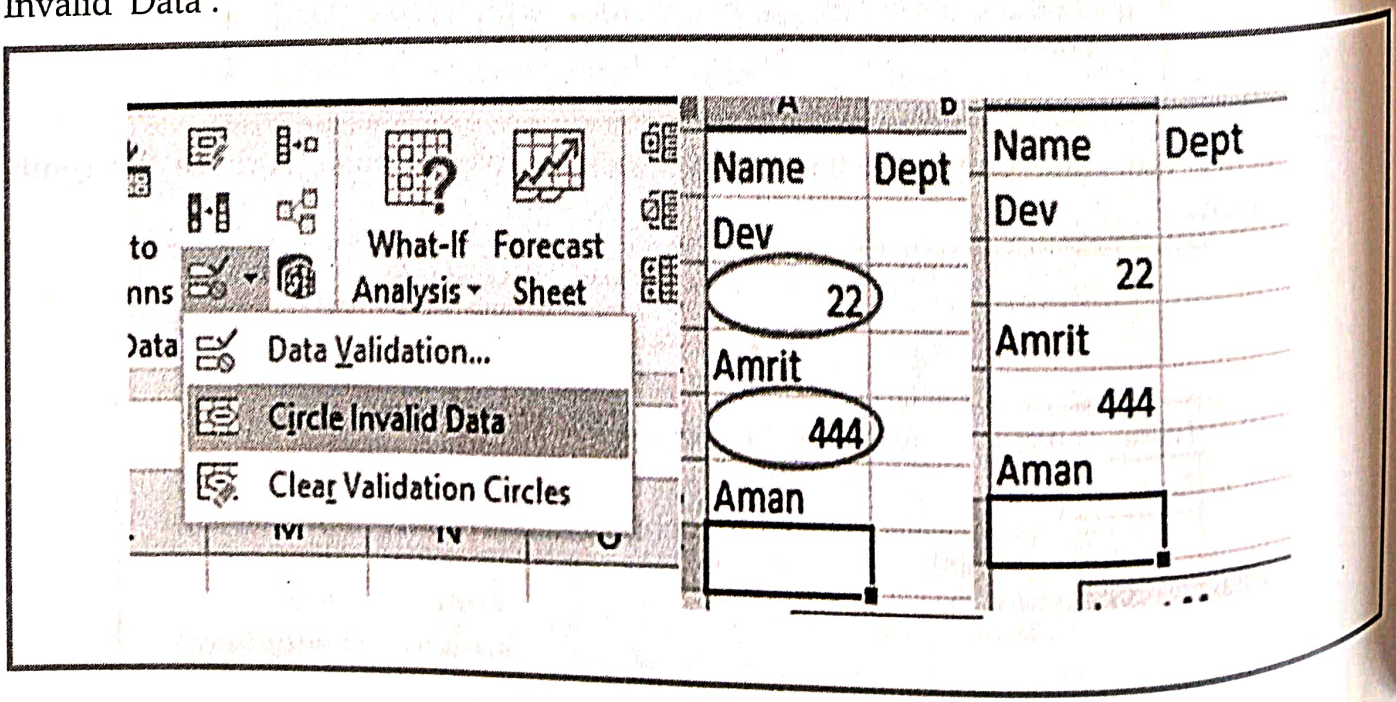
7.5.5 CIRCLE INVALID DATA

This option is useful if you don't the Stop error style or if you apply data validation to cells that already have values. When you choose this option and enter invalid data in the cell then circle appear around cells that contain incorrect entries.

If you have already applied the data validation to the cells which shows the error alert or applying the new validation on the cell then open the data validation window then click on Error Alter tab and uncheck the 'Show error alter after invalid data I entered' check box. Now enter the data as shown in second figure.



Now select the circle invalid data in Data Tab. It will circle all the invalid data cells. Clear the circle on the invalid cell using option 'Clear validation Circles, next to 'Circle Invalid Data'.



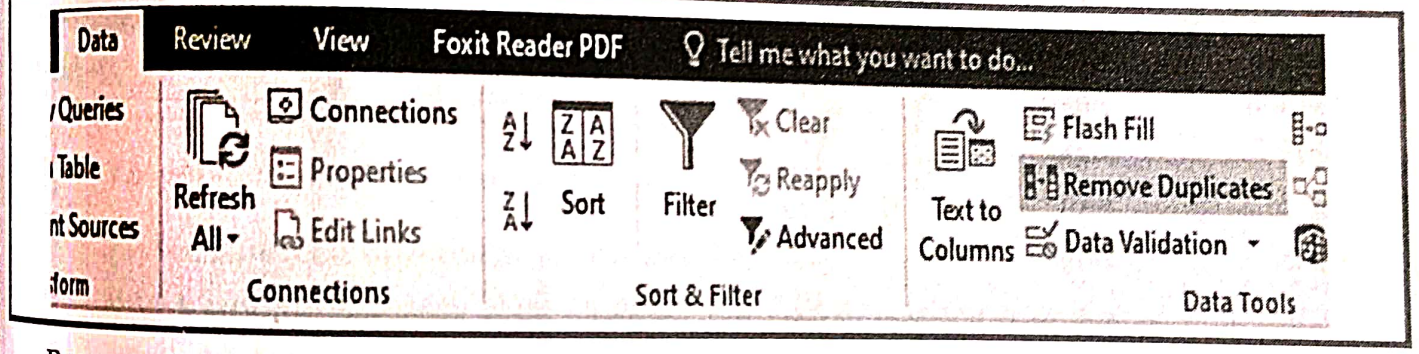
7.5.6 LIMITATION OF DATA VALIDATION

If you paste data into a restricted cell (after having copied it from somewhere) instead of typing it in, the data validation of the cell will be deleted. Neither the data you are pasting nor any future data entry into the cell will be checked anymore. Unfortunately, there is no simple way to guard against this problem.

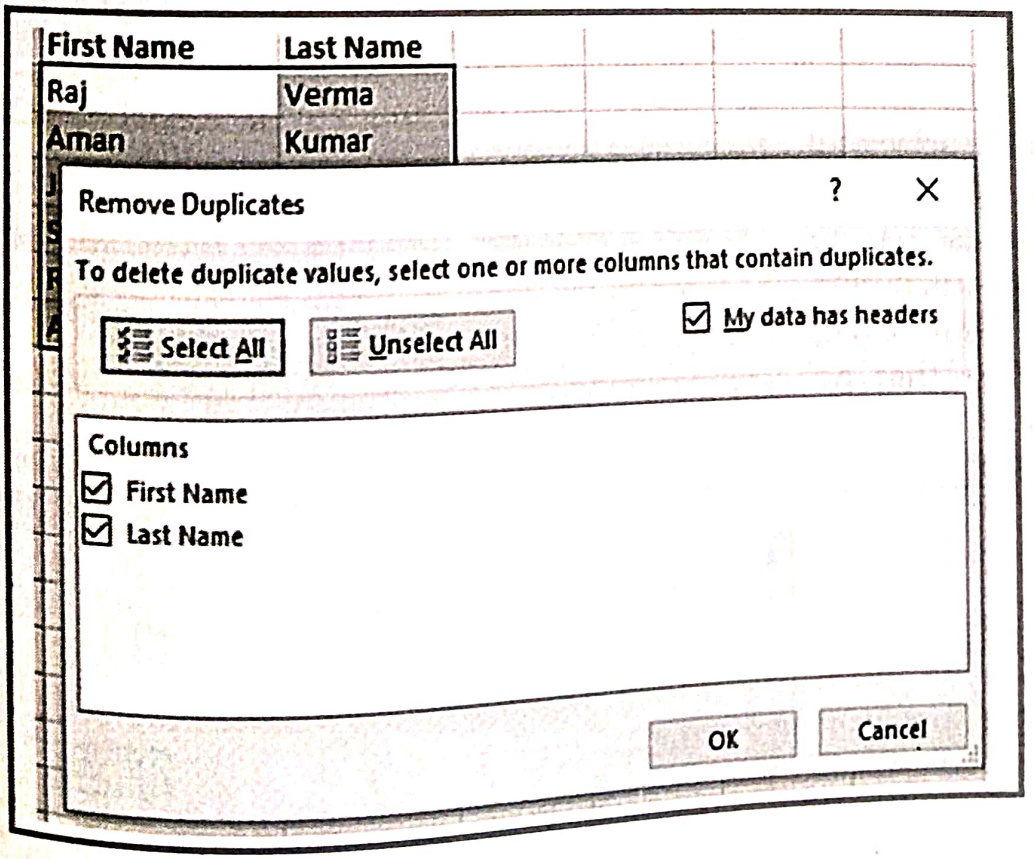
7.6 REMOVING DUPLICATE

A classic data entry problem, one particularly besetting large databases, is the spectre of duplicate records. It's not uncommon to discover the same names appearing repeatedly in large lists. Databases often include duplicate records that need to be deleted. When it has to be done it manually, it will take generally take a lot of time. Moreover, in a large database it will become more complicated. But Table makes it easy to delete duplicate records. Excel's tables are equipped with a Remove Duplicates feature, which speeds the task of winnowing those doubles from the data.

Start by moving the cell cursor to any cell within your data range. Click on Data Tab and Click Remove Duplicate option as shown below

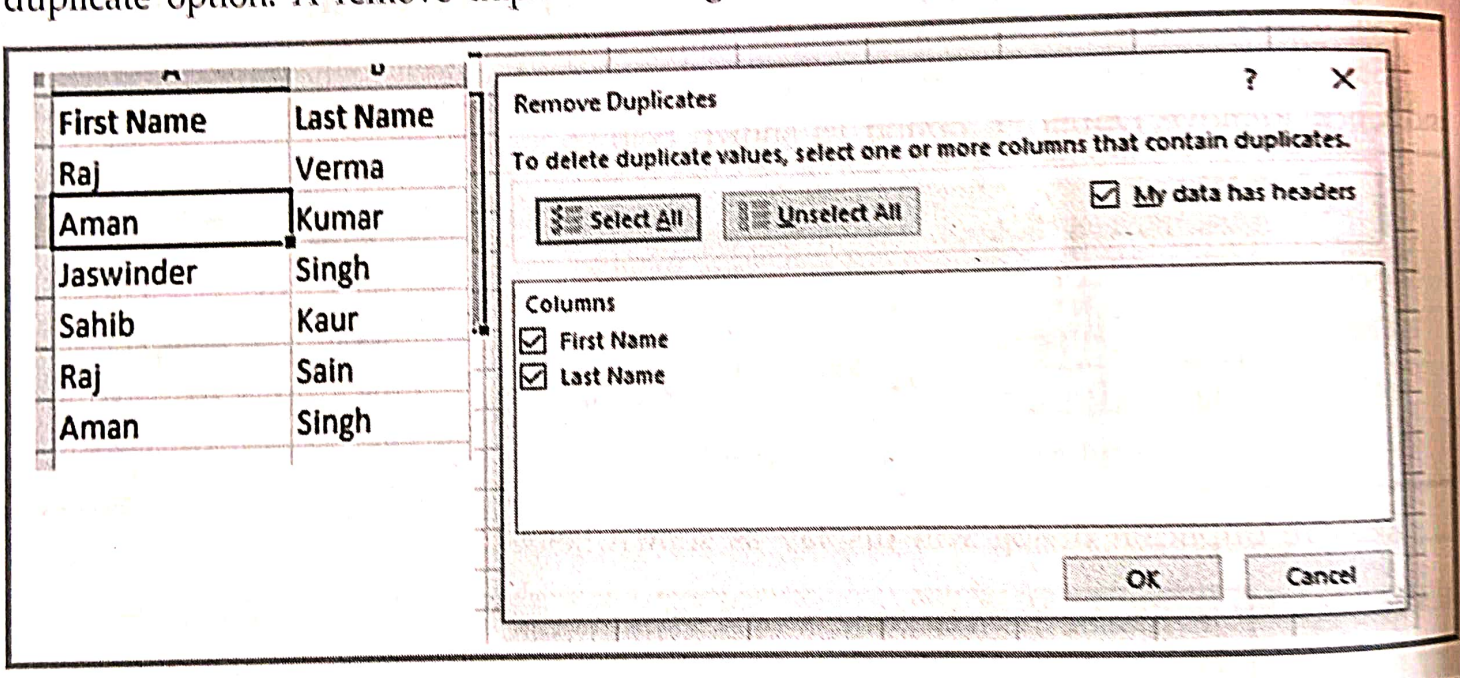


Remove duplicate dialog will display as shown below.

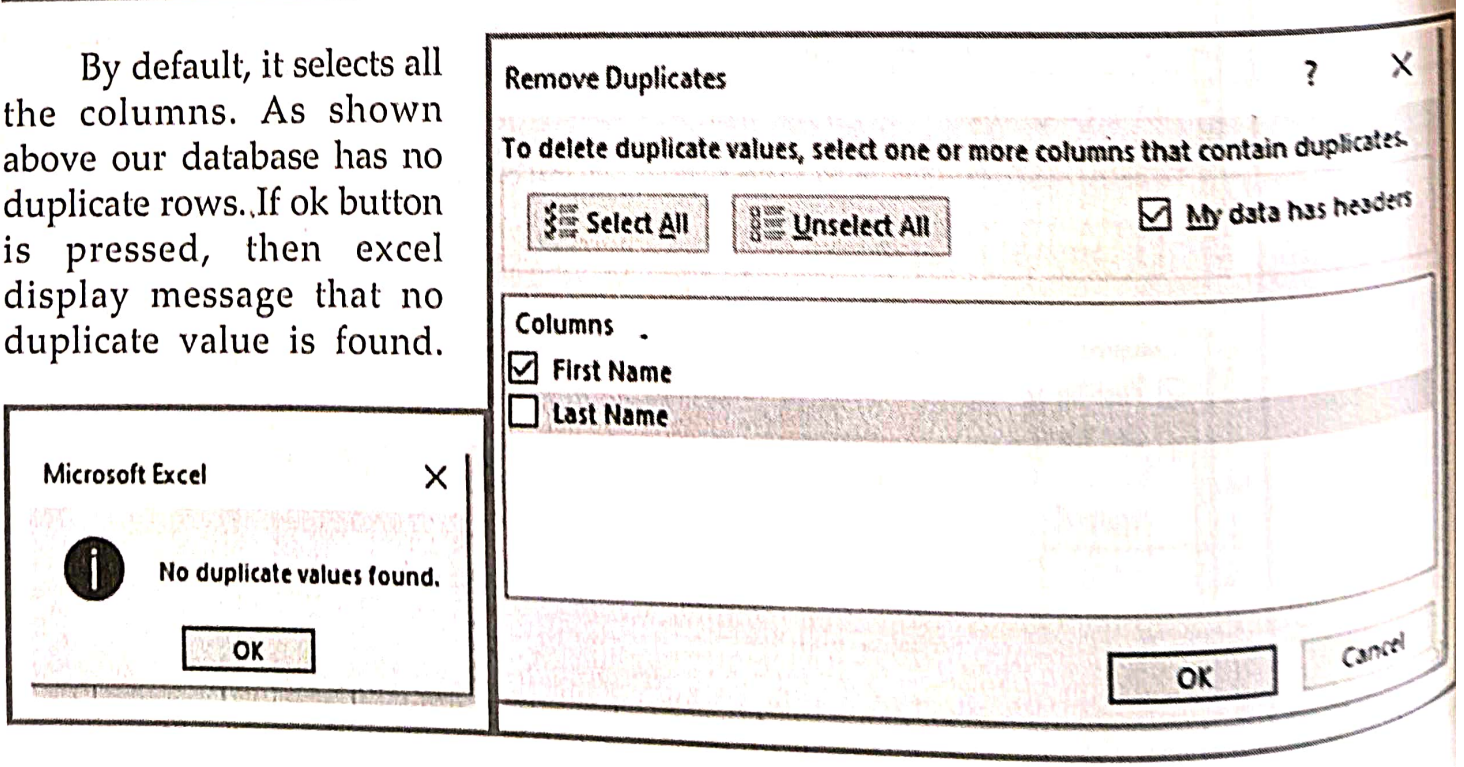


The Remove Duplicates dialog box lists all the columns in your data range or table. Place a check mark next to the columns that you want to be included in the duplicate search. Most of the time, you'll want to select all the columns, which is the default. Click OK, and Excel weeds out the duplicate rows and displays a message that tells you how many duplicates it removed. When you select all columns in the Remove Duplicates dialog box, Excel will delete a row only if the content of every column is duplicated. In some situations, you may not care about matching some columns, so you would deselect those columns in the Remove Duplicates dialog box. For example, if each row has a unique ID code, Excel would never find any duplicate rows. So you'd want to uncheck that column in the Remove Duplicates dialog box. When duplicate rows are found, the first row is kept and subsequent duplicate rows are deleted.

Let's take an example with the following small database which has some duplicate values in the column. Place the cursor in the data anywhere and then select remove duplicate option. A remove duplicate dialog box will appear.

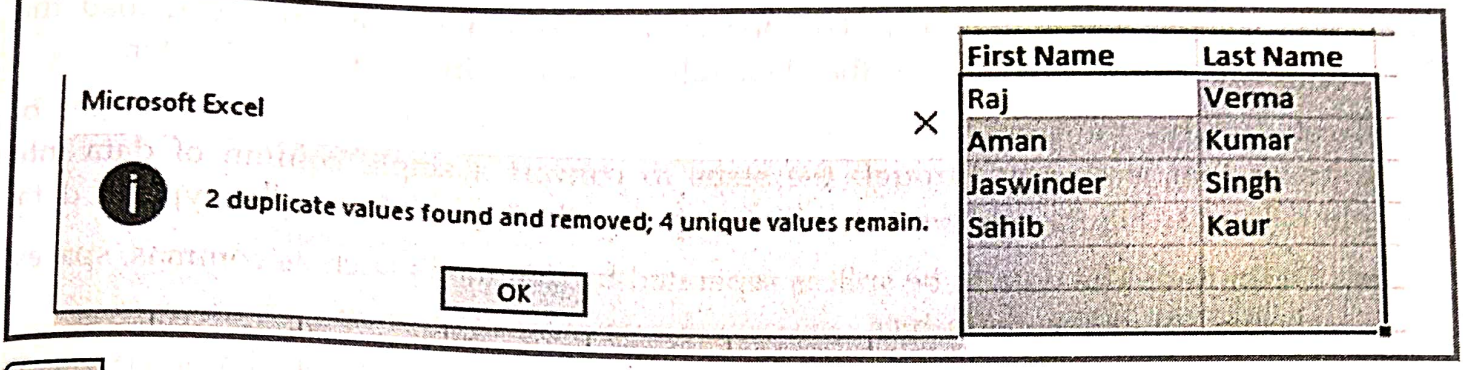


By default, it selects all the columns. As shown above our database has no duplicate rows. If ok button is pressed, then excel display message that no duplicate value is found.



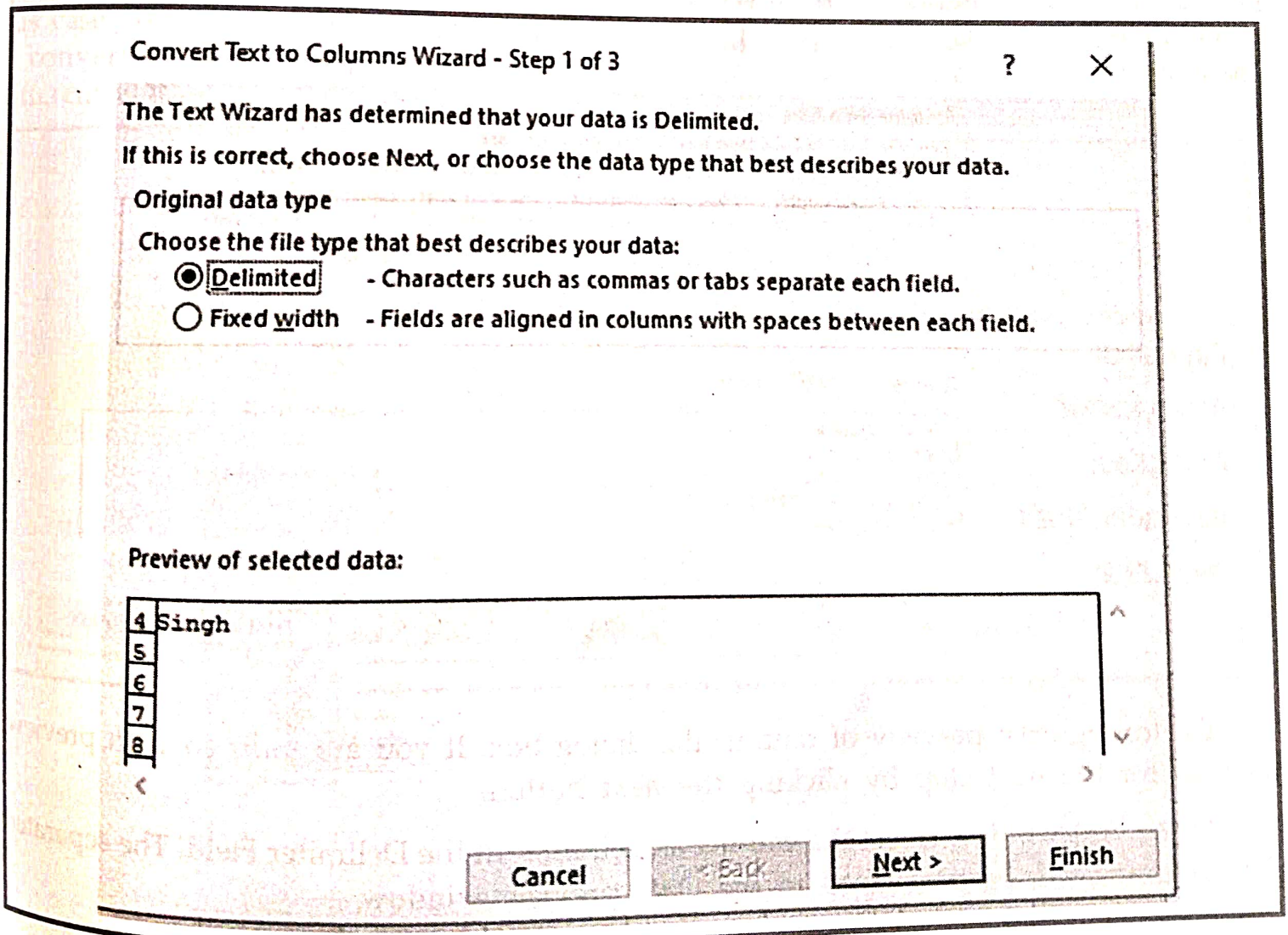
Now again press remove duplicate option and deselect the Last Name check box in the dialog box.

In our database 'First Name' column has two duplicate values ie. 'Raj' and 'Aman, Excel remain the first record of Raj and Aman and delete the second record when ok is pressed.



7.7 TEXT TO COLUMNS

Text to Columns is a great feature that lets you split your data into columns using a delimiter, which simply a text character that signifies the breaks between the different data points. If your delimiter is say, comma, Text to columns will put the data to the left of the comma in one column, the data to the right in another, and then it will delete the comma.



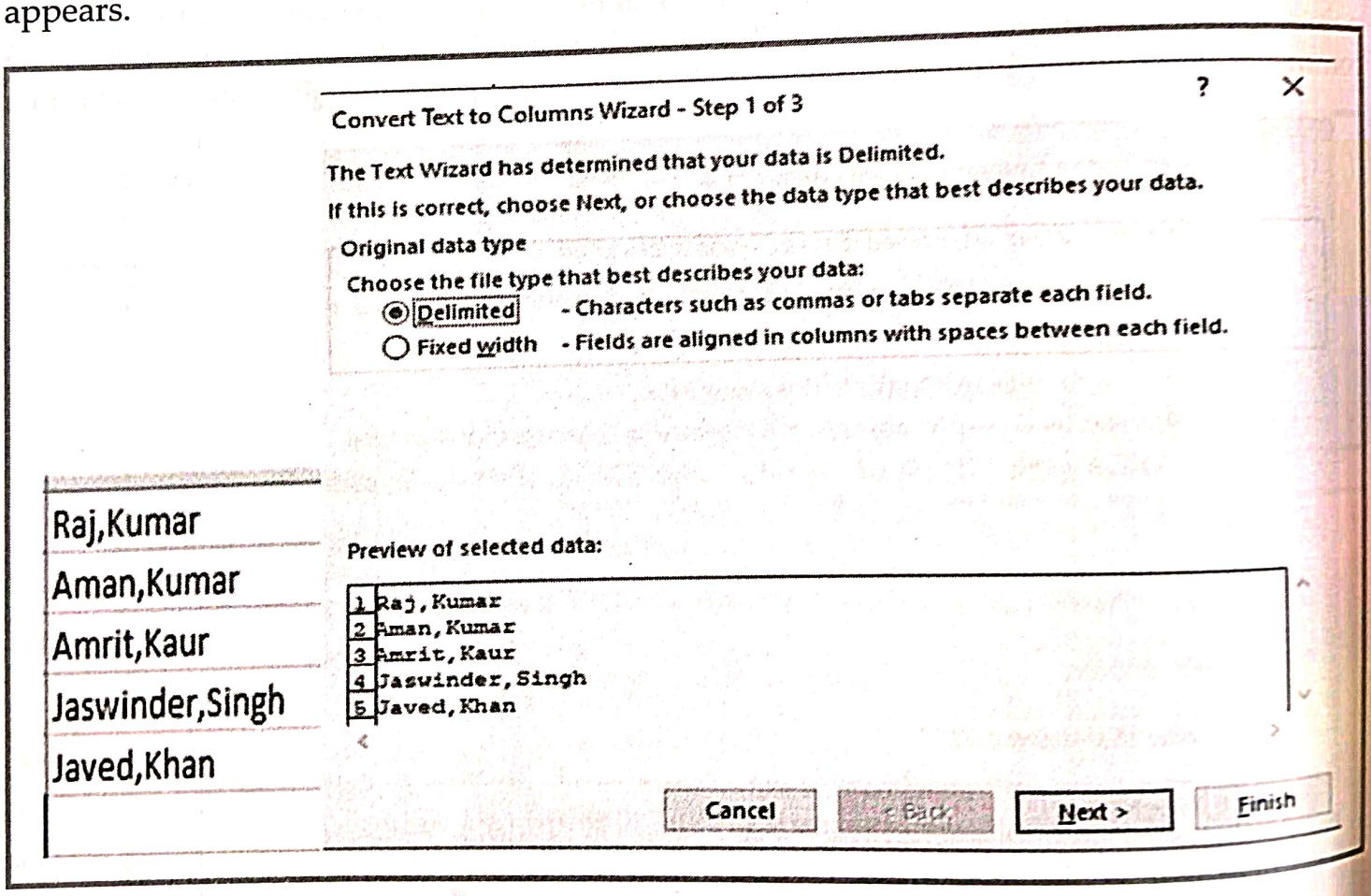
If you have more than one type of delimiters, you might have to run Text to Columns more than once. In this case, you have a period acting as delimiter as well as a comma and you could even those characters as delimiters, which would make Excel throw them out.

CSV is a popular format for data. The letters stand for Comma Separated Value. For these files, comma act as delimiter. This format is so common that when you load the CSV fie, Excel automatically splits the data into column using comma delimiter.

Excel displays the Convert Text to Columns Wizard, which consists of a series of dialog boxes that walk you through the steps to convert a single column of data into multiple columns. Figure shows the initial step, in which you choose the type of data:

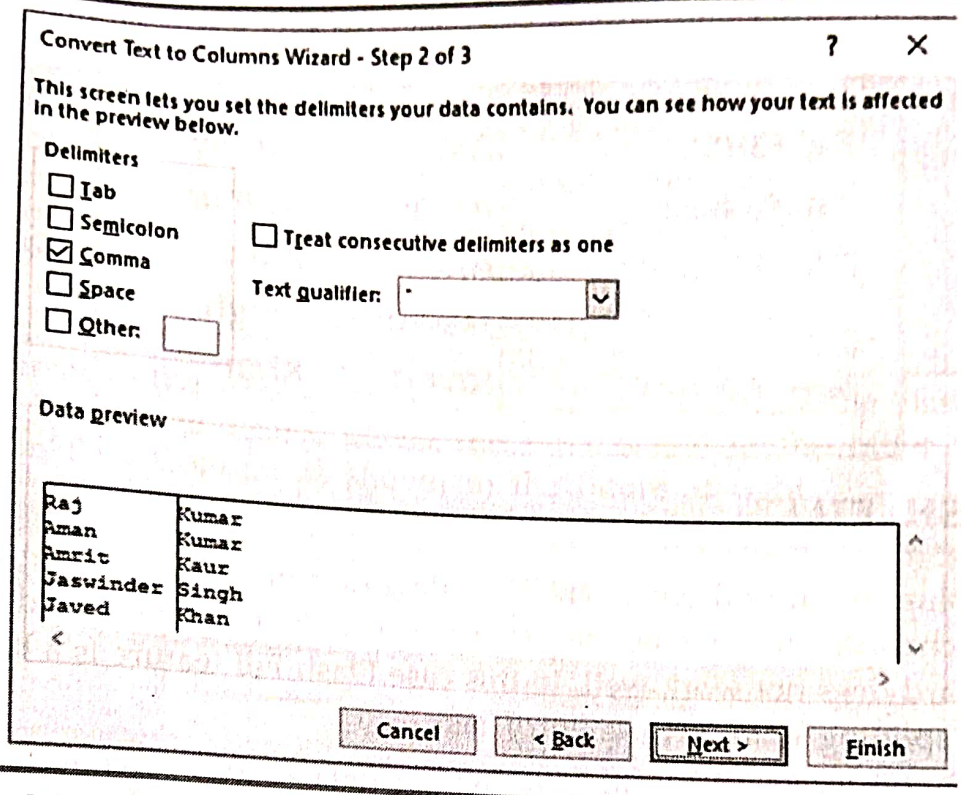
- **Delimited:** The data to be split is separated by delimiters such as commas, spaces, slashes, or other characters.
- **Fixed Width:** Each component occupies the same number of characters.

Consider the following database has small comma separated data. To convert into this data in two columns. Click on Text to Columns Options Data tab a dialog box appears.

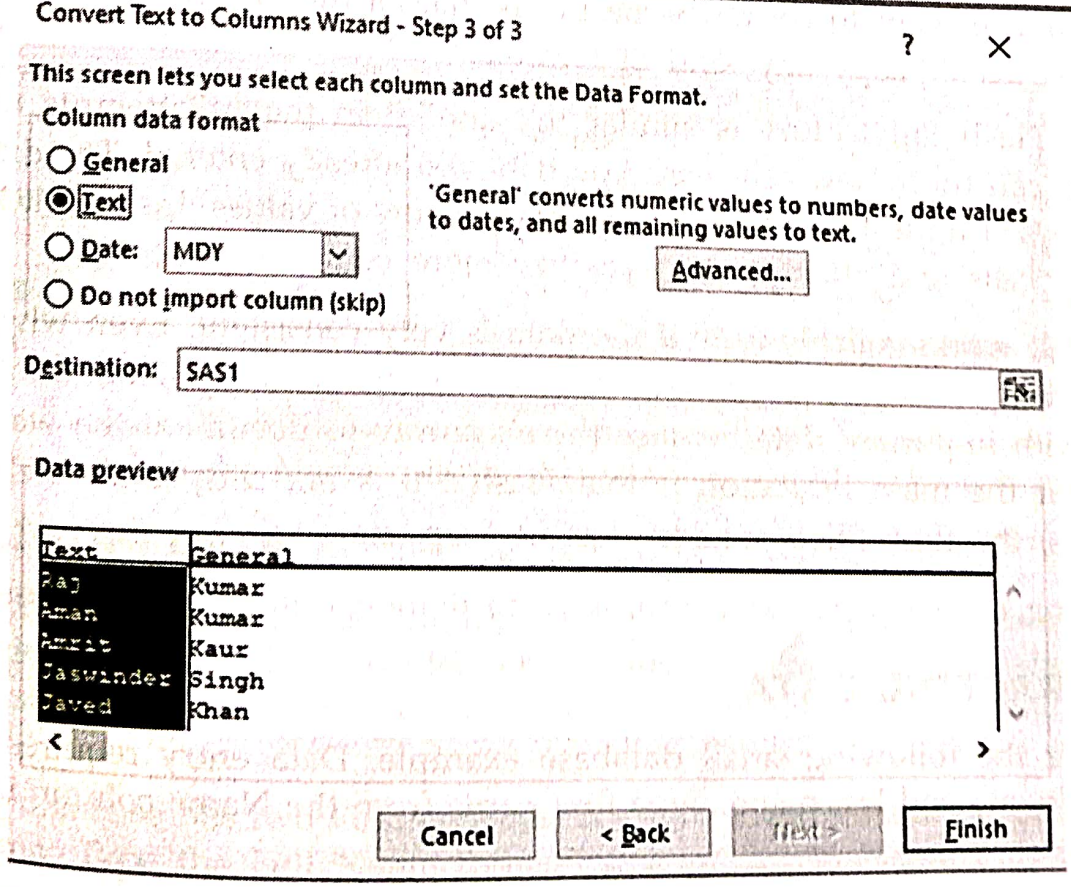


Go through the preview of data in the dialog box. If you are satisfied with preview, then go for the next step by clicking the next button.

In the next window select the comma as delimiter in the Delimiter Field. The separated data columns will be previewed in the Data preview window.



. Click on the Next button a next screen appears. Select the column data format as in this case 'Text' data format is selected. Select the destination field, choose whether data is converted in the same column or different columns. In our case we have choose different cell.



Click on finish button and Text is converted to different columns as shown below.

Raj,Kumar	Raj	Kumar
Aman,Kumar	Aman	Kumar
Amrit,Kaur	Amrit	Kaur
Jaswinder,Singh	Jaswinder	Singh
Javed,Khan	Javed	Khan

7.8 FLASH FILL

Text to Column is useful when data has delimiter or has fixed length. But data does not have both characteristics i.e. neither data has any delimiter nor fixed length then Text to Column wizard does not work well. In this case Flash Fill feature is a handy tool to help in cleaning or splitting data in data entry task.

Flash Fill uses pattern recognition to extract data (and also concatenate data). Flash Fill feature gives you the ability to take a part of the Data entered into one column of a worksheet table and enter just that data in a new table column using only a few keystrokes. The series of entries appear in the new column, literally in a flash (thus, the name Flash Fill), the moment Excel detects a pattern in your initial data entry that enables it to figure out the data you want to copy. The beauty is that all this happens without the need for you to construct or copy any kind of formula.

Excel's Flash Fill feature is similar to AutoFill in that it attempts to determine a pattern so it can fill in new cells based on data you already entered. The difference is that Flash Fill doesn't limit itself to a single column or row of values. Instead, it looks at all the surrounding cells and attempts to recognize more complex patterns.

Flash Fill works reliably only if the data is very consistent. Even when you think it worked correctly, make sure you examine the results carefully. And think twice before trusting it with important data because there's noway to document the way the data was extracted. But the main limitation is that Flash Fill is not a dynamic technique. If your data changes, the flash-filled column does not update.

Following are some working examples to demonstrate the usefulness of flash fill.

7.8.1 EXTRACTING DATA

Consider the following small database example. Data entry requires Name in two part i.e. first name and last name. Type first name from the Name column i.e. 'Raj' to First column.

	First	Last
Name		
Raj Kumar		
Aman Kumar		
Amrit Kaur		
Jaswinder Singh		
Javed Khan		

	A	B	C
1	Name	First	Last
2	Raj Kumar	Raj	
3	Aman Kumar		
4	Amrit Kaur		
5	Jaswinder Singh		
6	Javed Khan		

Move the cursor to the second row of First column. Click on the Flash Fill option in Data Tab. Flash Fill automatically detect the First Name in the other rows of Name column and fill in First Column as shown in the figure at right.

	A	B	C
1	Name	First	Last
2	Raj Kumar	Raj	
3	Aman Kumar		
4	Amrit Kaur		
5	Jaswinder Singh		
6	Javed Khan		

	A	B	C
1	Name	First	Last
2	Raj Kumar	Raj	
3	Aman Kumar	Aman	
4	Amrit Kaur	Amrit	
5	Jaswinder Singh	Jaswinder	
6	Javed Khan	Javed	

Now type the last name in the last column from the Name column and move the cursor to the next row of the Last Column and Use flash fill to fill the other rows.

	Name	First	Last
Raj Kumar	Raj	Kumar	
Aman Kumar	Aman		
Amrit Kaur	Amrit		
Jaswinder Singh	Jaswinder		
Javed Khan	Javed		

	A	B	C
1	Name	First	Last
2	Raj Kumar	Raj	Kumar
3	Aman Kumar	Aman	Kumar
4	Amrit Kaur	Amrit	Kaur
5	Jaswinder Singh	Jaswinder	Singh
6	Javed Khan	Javed	Khan

Now you can delete the Name column if it is not required or keep it for future use.

7.8.2 COMBINING DATA

If data is in different columns and we need to combine the data in a single column then flash fill do this job easily. Consider the pervious database example and we need to combine the first and last column in a single name column the following steps shows how it works.

Following database has two First and Last columns and we have to combine the data in Name column. Type the Full Name from First and Last to Name Column.

	A	B	C
1	First	Last	Name
2	Raj	Kumar	
3	Aman	Kumar	
4	Amrit	Kaur	
5	Jaswinder	Singh	
6	Javed	Khan	

	A	B	C
1	First	Last	Name
2	Raj	Kumar	Raj Kumar
3	Aman	Kumar	
4	Amrit	Kaur	
5	Jaswinder	Singh	
6	Javed	Khan	

Place the cursor in the second row of Name column and the use flash fill from Data tab. It will detect and fill the rest of the rows of the Name column automatically.

	A	B	C
1	First	Last	Name
2	Raj	Kumar	Raj Kumar
3	Aman	Kumar	
4	Amrit	Kaur	
5	Jaswinder	Singh	
6	Javed	Khan	

	A	B	C
1	First	Last	Name
2	Raj	Kumar	Raj Kumar
3	Aman	Kumar	Aman Kumar
4	Amrit	Kaur	Amrit Kaur
5	Jaswinder	Singh	Jaswinder Singh
6	Javed	Khan	Javed Khan

7.8.3 CHANGING CASE

Consider the data where Name column value in lower case and you want to Capitalize the first character of each word in the name. Flash Fill tool do this task in a second. Type the name in Sentence Case in New Name column. Then use Flash Fill to fill the other rows in the New Name column.

	A	B
1	Name	New Name
2	raj kumar	
3	aman kumar	
4	amrit kaur	
5	jaswinder singh	
6	javed khan	

	A	B
1	Name	New Name
2	raj kumar	Raj Kumar
3	aman kumar	
4	amrit kaur	
5	jaswinder singh	
6	javed khan	

	A	B
1	Name	New Name
2	raj kumar	Raj Kumar
3	aman kumar	Aman Kumar
4	amrit kaur	Amrit Kaur
5	jaswinder singh	Jaswinder Singh
6	javed khan	Javed Khan

7.8.4 ADDING TEXT

You can add the common text in the existing text using flash fill tool as shown below.

	A	B
1	Name	New Name
2	Raj Kumar	
3	Aman Kumar	
4	Amrit Kaur	
5	Jaswinder Singh	
6	Javed Khan	

	A	B
1	Name	New Name
2	Raj Kumar	Prof. Raj Kumar
3	Aman Kumar	
4	Amrit Kaur	
5	Jaswinder Singh	
6	Javed Khan	

A	B
Name	New Name
Raj Kumar	Prof. Raj Kumar
Aman Kumar	Prof. Aman Kumar
Amrit Kaur	Prof. Amrit Kaur
Jaswinder Singh	Prof. Jaswinder Singh
Javed Khan	Prof. Javed Khan

7.8.5 EXTRACTING PART FROM THE NUMBER

Flash fill also useful to extract part of the number. Following example demonstrate this concept.

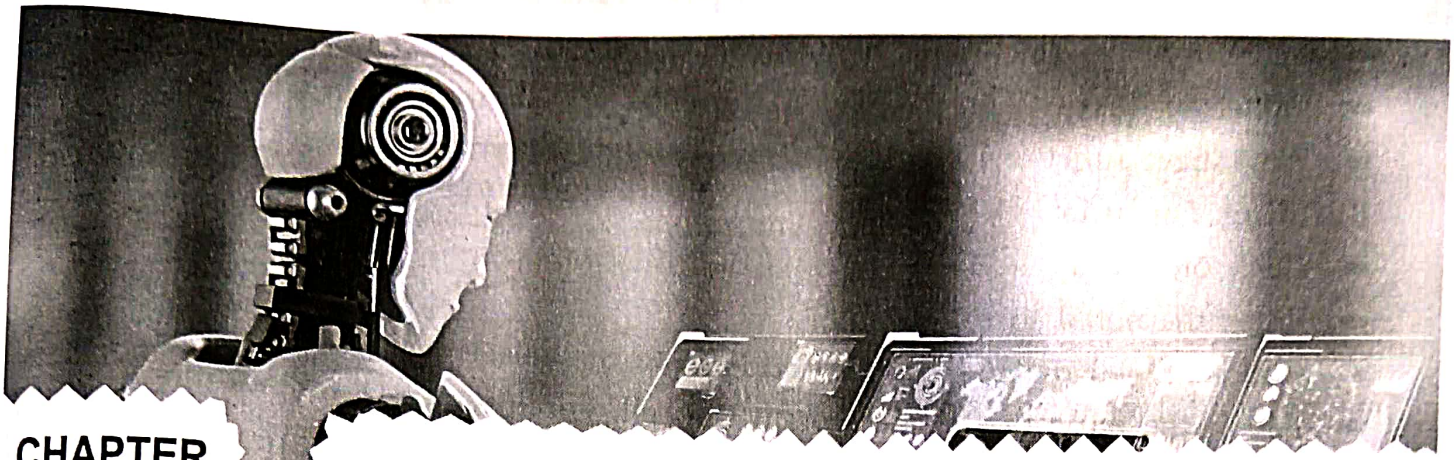
Type the whole price in the second column and use flash fill to do rest automatically.

	A	B
1	Price	Price in Whole Number
2	124.67	
3	786.94	
4	44.66	
5	889.98	
6	3456.88	

	A	D
1	Price	Price in Whole Number
2	124.67	124
3	786.94	
4	44.66	
5	889.98	
6	3456.88	

	A	D
1	Price	Price in Whole Number
2	124.67	124
3	786.94	786
4	44.66	44
5	889.98	889
6	3456.88	3456

☆☆☆☆☆



CHAPTER

8

MATH, STATISTIC AND TRIGONOMETRICEXCEL FUNCTION

CHAPTER AT A GLANCE

8.1 Intrduction	8.4 Statistical Functions
8.2 Basics of Formula and Function	8.5 Trigonometric Functions
8.3 Mathematical Function	

8.1 INTRODUCTION

Formulas are what make a spreadsheet program so useful. If it weren't for formulas, a spread- sheet would simply be a fancy word processing document that has great support for tabular information. You use formulas in your Excel worksheets to calculate results from the data stored in the worksheet. When data changes, the formulas calculate updated results with no extra effort on your part. This chapter introduces formulas and functions and helps you get up to speed with this important element.

8.2 BASICS OF FORMULA AND FUNCTION

8.2.1 FORMULA

A *formula* consists of special code entered into a cell. It performs a calculation of some type and returns a result that is displayed in the cell. Formulas use a variety of operators and worksheet functions to work with values and text. The values and text used in formulas can be located in other cells, which makes changing data easy and gives worksheets their dynamic nature.

A formula always begins with an equal sign and can contain any of these elements:

- Mathematical operators, such as + (for addition) and * (for multiplication)

- Cell references (including named cells and ranges)
- Values or Text
- Worksheet functions (such as SUM and COUNT)

Building Formula

- Click on a cell
- Enter the equal to (=) sign
- Choose a cell or enter the cell address in the selected cell
- Fill in the appropriate operator in the cell. For instance, + for addition.

	A	B	C
1			
2		Customers Sales Data	
3		Month	Sales
4		Jan	50000
5		Feb	75000
6		March	65000
7		Total Sale	=C4 + C5 + C6

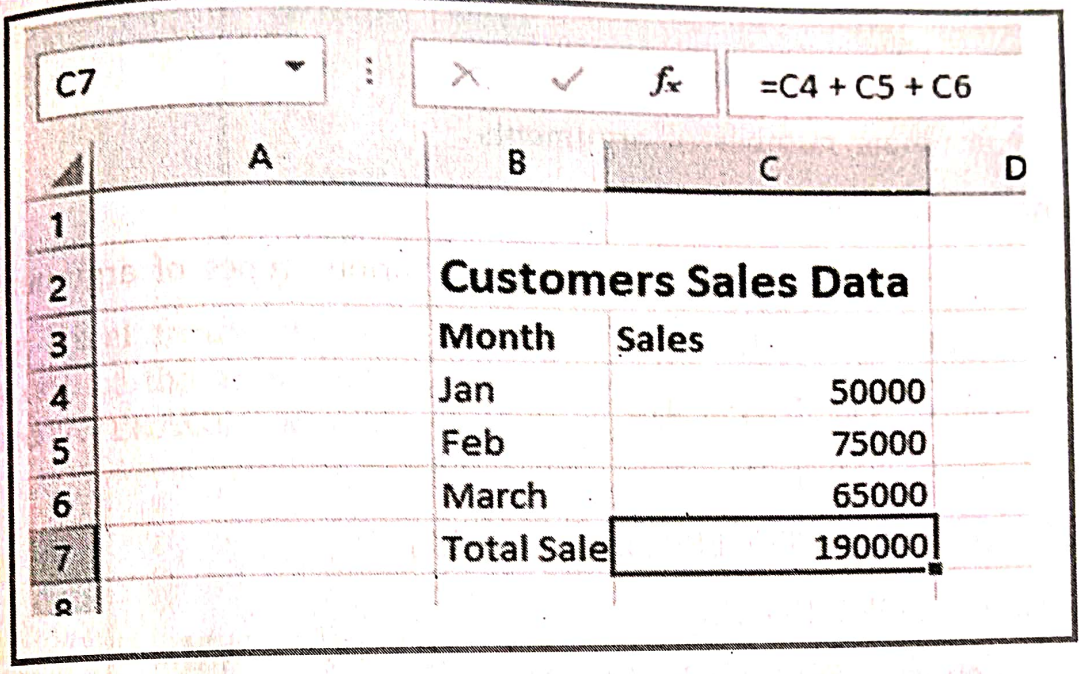
- Hit the Enter key and you will see the result of your calculation appearing inside the cell with the formula.

	A	B	C
1			
2		Customers Sales Data	
3		Month	Sales
4		Jan	50000
5		Feb	75000
6		March	65000
7		Total Sale	190000

Remember the rules of Brackets of Division Multiplication Addition & Subtraction (BODMAS). This means expressions are brackets are evaluated first. For arithmetic operators, the division is evaluated first followed by multiplication then addition and subtraction is the last one to be evaluated. For example, if formula is A2 * D2 / 2 then first it will divide D2 by 2 then mu. However, using bracket evaluation order can be changed. If above former formula is written as (A2 * D2)/2 then A2 and D2 will be multiplied first then divided by 2.

Checking a Formula: When you enter a formula inside a cell, you will also see the formula appearing in the formula bar. The space highlighted in red below is the formula bar.

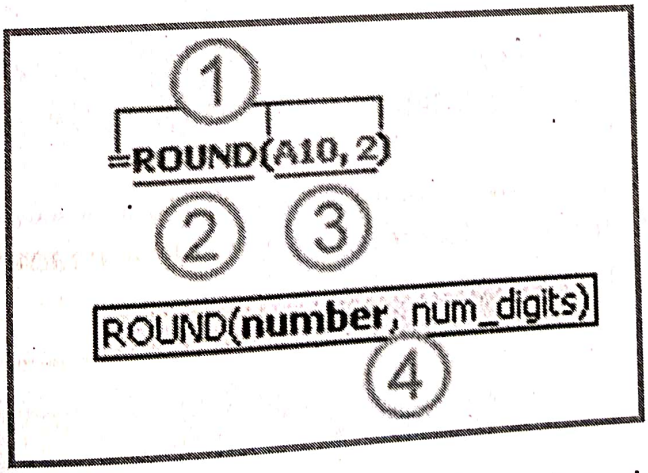
To view a particular formula, choose a cell containing the identity of the formula and you will see the formula appearing in the formula bar.



8.2.2 FUNCTION

A function is a preset formula in Excel. It allows you to build more complicated calculations without having to work out the algebra behind it.

Excel Function Syntax: To understand function syntax, consider the Round function below designed to round off a particular number in cell A10;



1 & 2 Structure: An equal to symbol (=) is used to initiate the structure of any function in Excel, then followed by the name of the function, opening parenthesis, the function's argument which is usually separated by a comma, and a closing parenthesis.

Note: You can insert function using insert function wizard discussed in next section.

3. Arguments: The arguments can be texts, numbers, logical values like **TRUE** or FALSE, error values like #N/A, arrays or cell references. The designated argument must be one that can give a value that is valid for the argument. The Arguments can as well be formulas, constant or some other functions.

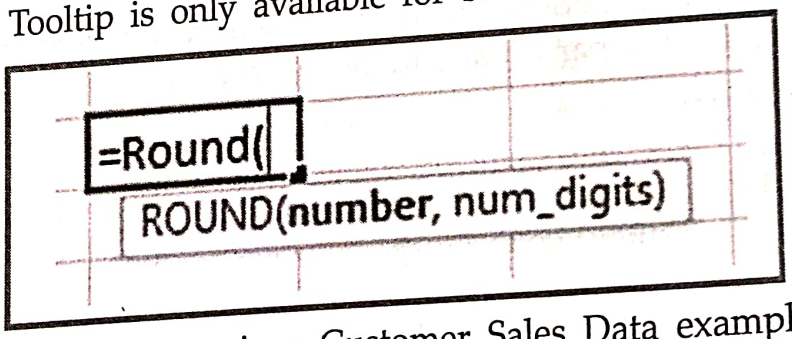
Functions vary in the way they use arguments. Depending on what it has to do, a function may use

- No arguments
- One argument
- A fixed number of arguments
- An indeterminate number of arguments
- Optional arguments

Here are some examples of functions that use various types of arguments:

- No argument: =TODAY()
- Cell reference: =SUM(A1:A24)
- Literal value: =SQRT(121)
- Literal text string: =PROPER("john f. smith")
- Expression: =SQRT(183+12)
- Other functions: =SQRT(SUM(A1:A24))

4. Argument tooltip: As you are entering the function, you will see a tooltip with the arguments and syntax. For instance, enter **=ROUND** (and you can see the tooltip coming up. Note that the Tooltip is only available for built-in functions.



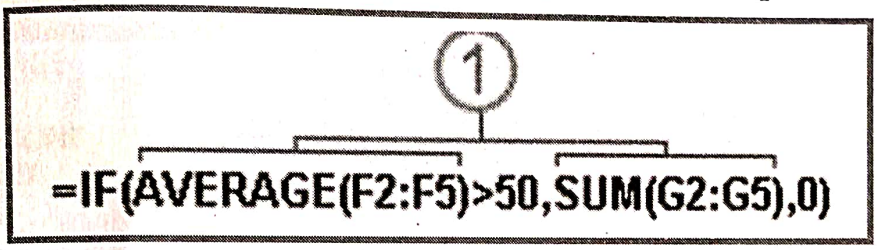
Let's continue with the previous Customer Sales Data example. This time calculate the total using SUM function.

	A	B	C
1			
2		**Customers Sales Data**	
3		**Month**	**Sales**
4		Jan	50000
5		Feb	75000
6		March	65000
7		Total Sale	
8			

	A	B	C
1			
2		**Customers Sales Data**	
3		**Month**	**Sales**
4		Jan	50000
5		Feb	75000
6		March	65000
7		Total Sale	=SUM(C4:C6)

B	C
Customers Sales Data	
Month	**Sales**
Jan	50000
Feb	75000
March	65000
Total Sale	190000

Nesting Excel functions: In few cases, there might be a need for you to use a function as one of the arguments of another function. For instance, the formula below uses a nested AVERAGE function and then compares the output with the value 50.



1. The SUM and the AVERAGE functions are nested within the IF function.

Valid Returns: When you use a nested function as an argument, the nested function must be able to return the exact type of values that the argument uses. For instance, the nested function must send back a TRUE or FALSE value. If the function is unable to do this, then Excel will bring a #VALUE! error value.

Nesting level limits: One formula can have up to 7 levels of nested functions. When a particular function, say function B, is deployed, as an argument in another function say Function A, we will say that the function B is acting as a second level function. For instance, the SUM function and the AVERAGE Function are considered to be second-level functions if they are deployed as arguments of the IF function. A function that is nested within the nested AVERAGE function will now be considered as a third level function, and so on

Note: Formula is broader term where arithmetic expression, constants, name, function or more than one of these terms are used. Functions may be the part of formula.

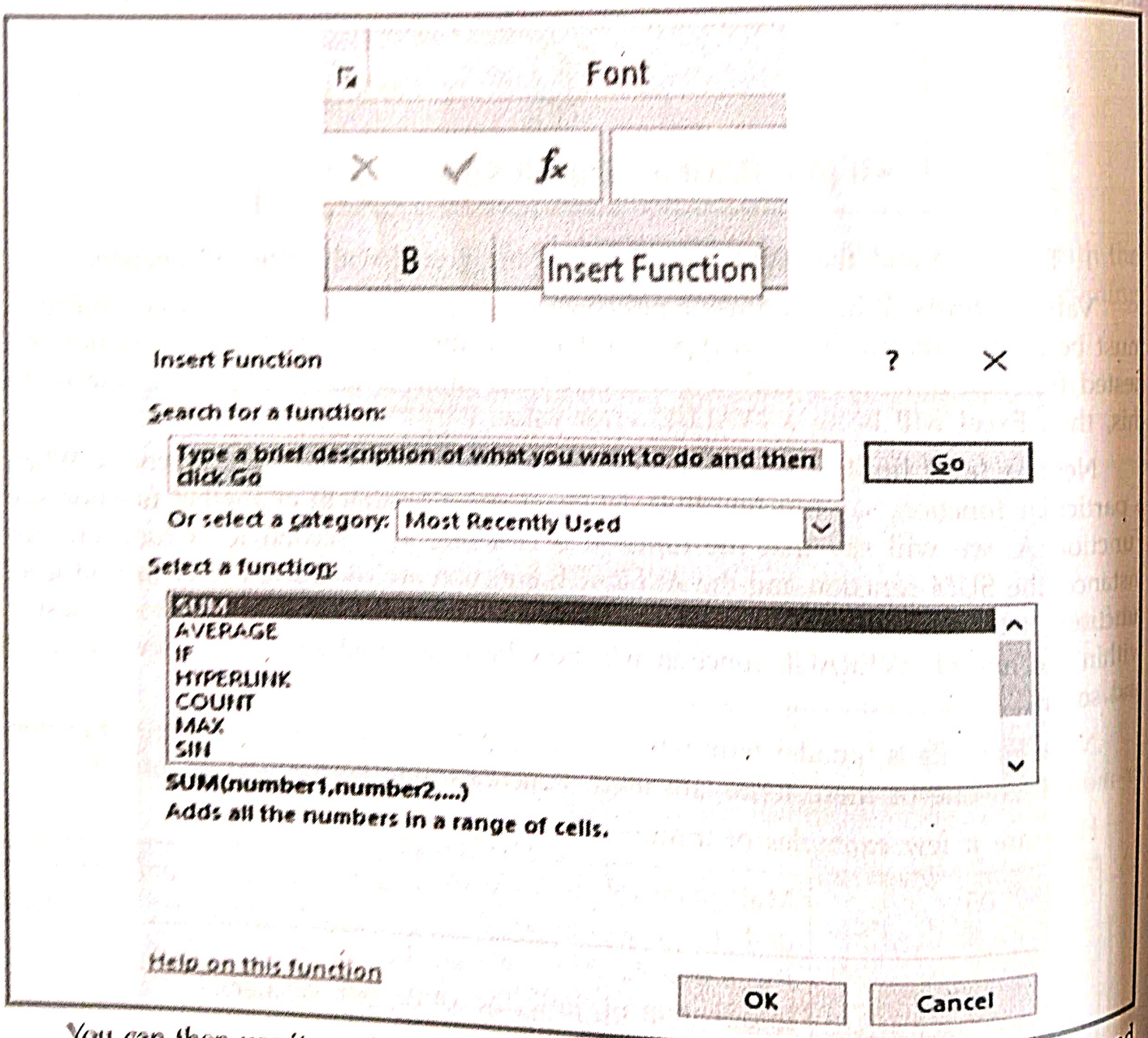
Here are a few examples of formulas :

= 150*.05	Multiplies 150 times 0.05. This formula uses only values, and it always returns the same result. You could just enter the value 7.5 into the cell, but using a formula provides information on how the value was calculated.
=A3	Displays the value in cell A3. No calculation is performed.
=A1 + A2	Adds the values in cells A1 and A2.

=Income-Expenses	Subtracts the value in the cell named Expenses from the value in the cell named Income.
= SUM(A1 : A12)	Adds the values in the range A1:A12, using the SUM function.
= A1 = C12	Compares cell A1 with cell C12. If the cells are identical, the formula returns TRUE, otherwise, it returns FALSE.

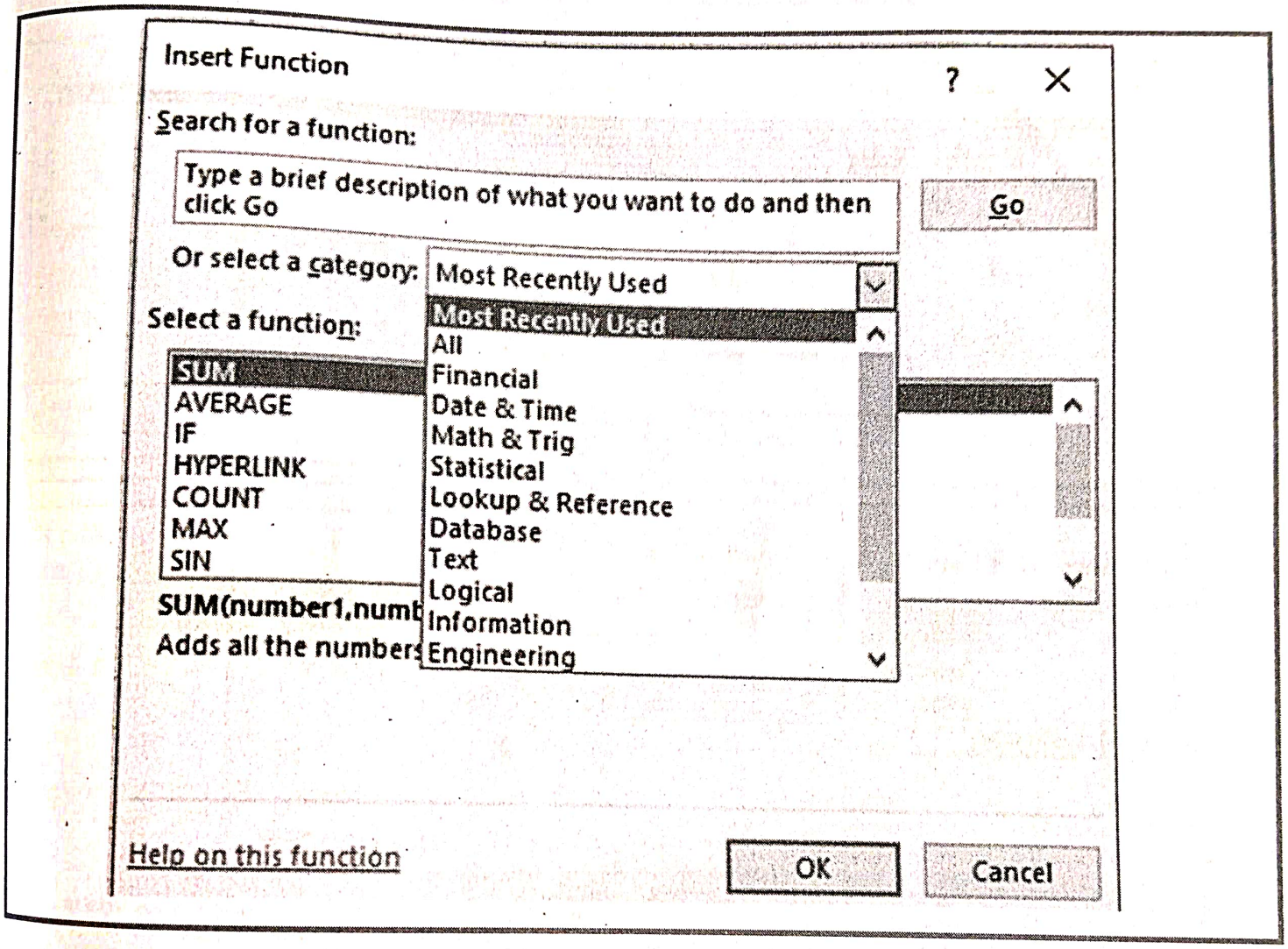
8.2.3 USING INSERT FUNCTION WIZARD

Excel supports a wide variety of built-in functions that you can use when building formulas. If you are unsure on which formula to use in Excel. You can use the *Insert Function Wizard of Excel* to find one for your purpose. When you click the Insert Function button, Excel displays the Insert Function dialogue box, similar to the one shown in Figure.



You can then use its options to find and select the function that you want to use and to define the argument or arguments that the function requires in order to perform its calculation.

You can select the most commonly used types of Excel functions and enter them simply by choosing their names from the drop-down menus attached to their command buttons in the Function Library group of the Formulas tab of the Ribbon. These command buttons include Financial, Logical, Text, Date & Time, Lookup & Reference, and Math & Trig.

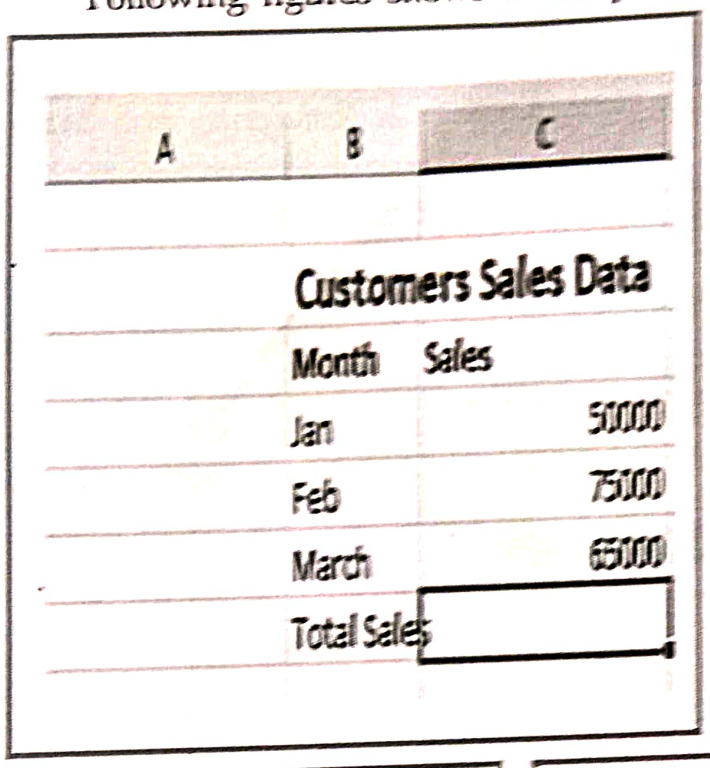


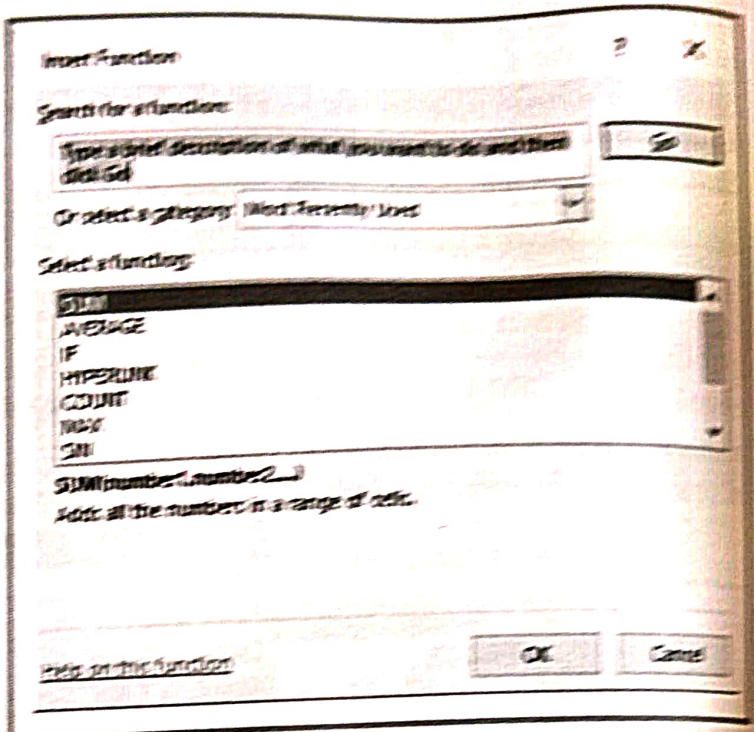
In addition, you can select functions in the Statistical, Engineering, Cube, Information, Compatibility, and Web categories from continuation menus that appear when you click the More Functions command button on the Formulas tab. And if you find you need to insert a function in the worksheet that you recently entered into the worksheet, chances are good that when you click the Recently Used command button, that function will be listed on its dropdown menu for you to select.

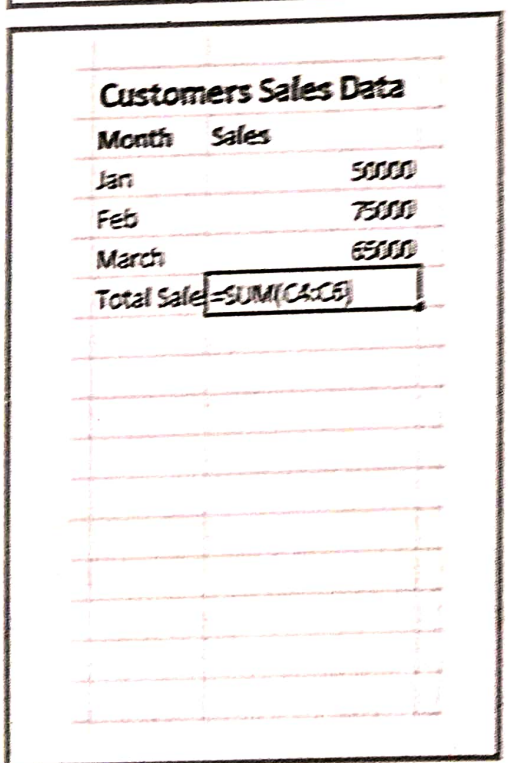
Let's continue with the previous Customer Sales Data example. This time calculate the total using SUM function instead of arithmetic expression.

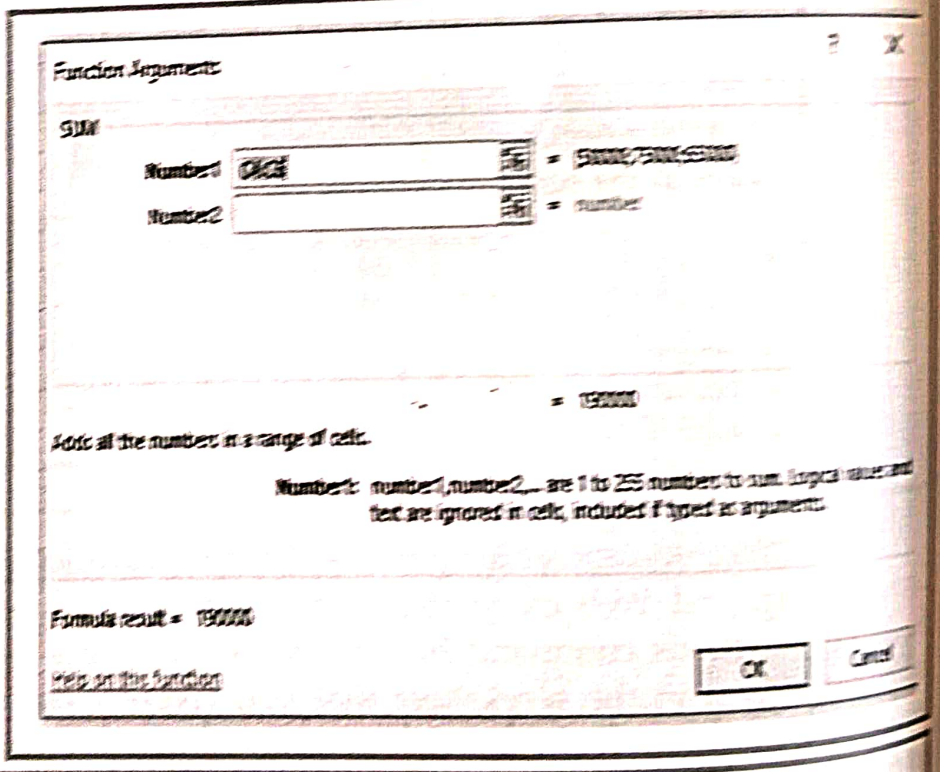
1. Select the cell where result has to be displayed.
2. Click on insert function button. An insert function wizard will be displayed.
3. Select SUM function and click ok button.
4. Enter range as an arguments of SUM function and click ok
5. Total will be displayed in selected cell.

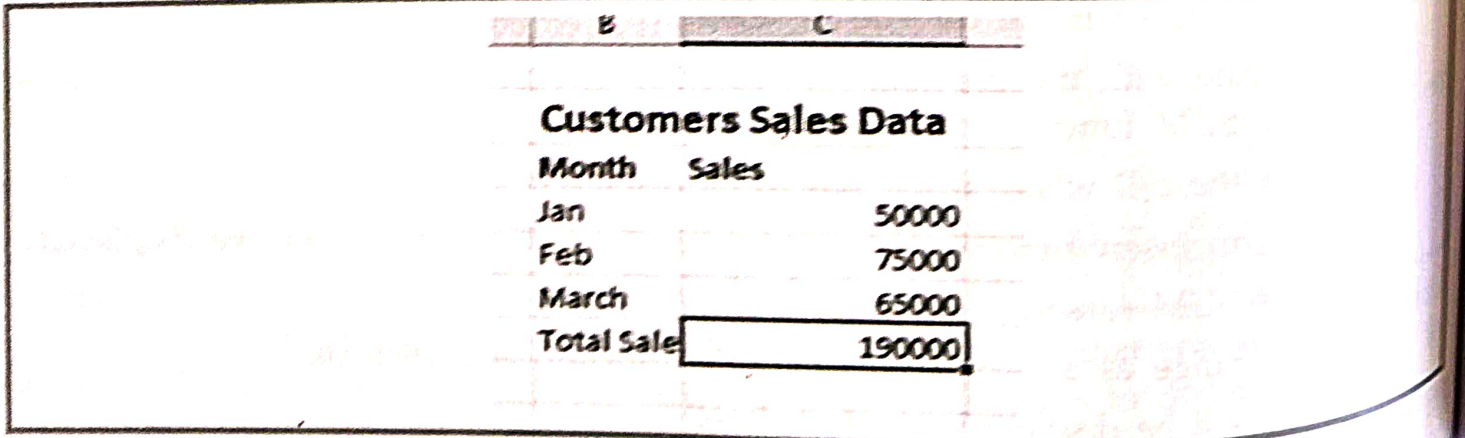
Following figures shows the steps discussed above.











8.2.4 OPERATORS IN EXCEL

Many of the simpler formulas that you build require the sole use of Excel's operators, which are the symbols that indicate the type of computation that is to take place between the cells and/or constants interspersed between them. Excel uses four different types of operators: arithmetic, comparison, text, and reference. Following table shows all these operators arranged by type and accompanied by an example.

Type	Character	Operation	Example
Arithmetic	+ (plus sign)	Addition	=A2+B3
	–(minus sign)	Subtraction or negation	=A3–A2 or –C4
	* (asterisk)	Multiplication	=A2*B3
	/	Division	=B3/A2
	%	Percent (dividing by 100)	= B3%
	^	Exponentiation	=A2^3
Comparison	=	Equal to	=A2=B3
	>	Greater than	=B3>A2
	<	Less than	=A2<B3
	>=	Greater than or equal to	=B3>=A2
	<=	Less than or equal to	=A2<=B3
	<>	Not equal to	=A2<>B3
Text	&	Concatenates (connects) entries to produce one continuous entry	=A2& " " &B3t
Reference	: (colon)	Range operator that includes	=SUM(C4:D17)
	, (comma)	Union operator that combines multiple references into one reference	=SUM(A2, C4:D17, B3)
	(space)	Intersection operator that produces one reference to cells in common with two references	=SUM(C3:C6 C3:E6)

Concatenation Operator: The ampersand symbol (&) is a concatenation operator that can be used to combine one or more text strings to get a single piece.

Text operator	Meaning	Example
& (ampersand)	Connects, or connactenates, two values to produce one continuous text value	='North"&" wind" results in 'Northwind", Where A1 holds "Last name" and B1 holds "First name", = A1&", " &B1 results in "Last name, First Name".

Order of operator precedence: When you build a formula that combines different operators, Excel follows the set order of operator precedence, as shown in Table. When you use operators that share the same level of precedence, Excel evaluates each element in the equation by using a strictly left-to-right order.

Precedence	Operator	Type/Function
1	–	Negation
2	%	Percent
3	^	Exponentiation
4	*and/	Multiplication and Division
5	+ and –	Addition and Subtraction
6	&	Concatenation
7	=, <, >, <=, >=, <>	All Comparison Operators

8.3 MATHEMATICAL FUNCTION

Excel's mathematical underpinnings are revealed in the long list of math-related functions that come with the program. Functions exist for basic mathematical operations such as absolute values, lowest and greatest common denominators, square roots, and sums. Plenty of high-end operations also are available for things such as matrix multiplication, multinomial, and sums of squares. Not all of Excel's math functions are useful in a business context, but a surprising number of them are. For example, operations such as rounding and generating random numbers have business uses. This chapter introduces you to these and many other math-related worksheet functions.

8.3.1. ABS FUNCTION

ABS(number)

The ABS function returns the absolute value of a number—that is, the number without its sign. With this function, the argument **number is** the real number of which you want the absolute number.

	A	B	C	D
1	Location	Expected Change in Sales	Abs. Change in Sales	
2	A	-8%	8.00%	=ABS(B2)
3	B	5%	5.00%	
4	C	10%	10.00%	
5	D	-12%	12.00%	

8.3.2. FACT FUNCTION

FACT(number)

The FACT function returns the factorial of a number. The factorial of a number is equal to 1 × 2 × 3 ×... × number. number is the non negative number of which you want the factorial. If number is not an integer, it is truncated. By definition, FACT(0) is 1.

B11 =FACT(A11)

	A	B	C
1	Number	Factorial	Equivalent to
2	0	1	
3	1	1	=1
4	2	2	=1*2
5	3	6	=1*2*3
6	4	24	=1*2*3*4
7	5	120	=1*2*3*4*5
8	6	720	=1*2*3*4*5*6
9	7	5040	=1*2*3*4*5*6*7
10	8	40320	=1*2*3*4*5*6*7*8
11	9	362880	=1*2*3*4*5*6*7*8*9
12	99	9.3E+155	=1*2*3*...*98*99

8.3.3. GCD AND LCM FUNCTION

The GCD function returns the greatest common divisor of two or more integers. The greatest common divisor is the largest integer that divides both number 1 and number 1 without a remainder.

GCD(number 1, number 2, ...)

- **number1, number2, ...** are 1 to 29 values. If any value is not an integer, it is truncated. If any argument is non numeric, GCD returns a #VALUE! error. If any argument is less than zero, GCD returns a #NUM! error.

The number 1 divides any value evenly. A prime number has only itself and 1 as even divisors.

116	Number 1	Number 2	Number 3	GCD	
117	7	3	12	7	=GCD(A117)
118	7	3	12	1	=GCD(A118:B118)
119	7	3	12	1	=GCD(A119:C119)

The LCM function returns the least common multiple of integers. The least common multiple is the smallest positive integer that is a multiple of all integer arguments—number1, number2, and so on. You use LCM to add fractions with different denominators.

LCM (number 1, number2, ...)

- number1, number2, ... are 1 to 29 values for which you want the least common multiple. If the value is not an integer, it is truncated. If any argument is non numeric, LCM returns a #VALUE! error. If any argument is less than one, LCM returns a #NUM! error

116	Number 1	Number 2	Number 3	LCM	
117	7	3	12	7	=LCM(A117)
118	7	3	12	21	=LCM(A118:B118)
119	7	3	12	84	=LCM(A119:C119)

8.3.4. MOD AND QUOTIENT FUNCTION

The MOD function returns the remainder after number is divided by divisor. The result has the same sign as divisor.

MOD (number, divisor)

This function takes the following arguments:

- number This is the number for which you want to find the remainder.
- divisor This is the number by which you want to divide number. If divisor is 0, MOD returns a #DIV/0! error.

The MOD function is good for classifying records that follow a certain order.

134	Number	Divisor	MOD	
135	5	2	1	=MOD(A135,B135)
136	5	-2	-1	=MOD(A136,B136)
137	-5	2	1	=MOD(A137,B137)

The QUOTIENT function returns the integer portion in a division problem. You use this function when you want to discard the remainder in a division problem.

QUOTIENT(numerator, denominator)

This function takes the following arguments:

- numerator This is the dividend.
- denominator This is the divisor.

If either argument is non numeric, QUOTIENT returns a #VALUE! error

1	Number	Divisor	Quotient
2	1	4	0
3	2	4	0
4	3	4	0
5	4	4	1
6	5	4	1
7	6	4	1
8	7	4	1
9	8	4	2
10	9	4	2
11	10	4	2

8.3.5. SQRT AND POWER

SQRT(number)

The SQRT function returns a positive square root. The argument number is the number for which you want the square root. If number is negative, SQRT returns a #NUM! error.

208	Value	Square Root	
209	225		15 =SQRT(A209)
210	-9	#NUM!	=SQRT(A210)
211	0		0 =SQRT(A211)

POWER (number, power)

The POWER function returns the result of a number raised to a power. This function takes the following arguments:

- **number** This is the base number. It can be any real number.
- **power** This is the exponent to which the base number is raised.

The POWER function works with all sorts of irrational numbers, such as 98.2 raised to the 3.4 power.

140	Number	Exponent	POWER	
141	2	3	8	=POWER(A141,B141)
142	2	4	16	=POWER(A142,B142)
143	2	-2	0.25	=POWER(A143,B143)

8.3.6. ROMAN FUNCTION

The ROMAN function converts an Arabic numeral to Roman, as text.

ROMAN (number, form)

This function takes the following arguments:

- **number** This is the Arabic numeral you want converted.
- **form** This is a number that specifies the type of Roman numeral you want. The Roman numeral style ranges from Classic to Simplified, becoming more concise as the value of form increases.

3	MMVII	=ROMAN(2007)
4	MMVII	=ROMAN(YEAR(NOW()))
5		
6	MCMXCIX	=ROMAN(1999)
7	MCMXCIX	=ROMAN(1999,0)
8	MLMVLIV	=ROMAN(1999,1)
9	MXMIX	=ROMAN(1999,2)
10	MVMIV	=ROMAN(1999,3)
11	MIM	=ROMAN(1999,4)

8.3.7. ROUND, MROUND, ROUNDUP, ROUNDDOWN FUNCTIONS

The most basic (and most commonly used) of Excel's rounding functions is ROUND(), which rounds a numeric value to whatever level of precision you choose.

ROUND (number, num_digits)

The ROUNDDOWN() and ROUNDUP() functions are similar to ROUND(), except that they always round in a single direction: ROUNDDOWN() always rounds a number toward 0, and ROUNDUP()always rounds away from 0. Here is the syntax for each of these functions:

ROUNDDOWN(number, num_digits)
ROUNDUP(number, num_digits)

- **number** The number you want to round.
- **num_digits** An integer that specifies the number of digits you want number rounded to, as explained here: num_digits Description

  > 0 Rounds number to num_digits decimal places

  0 Rounds number to the nearest integer

  < 0 Rounds number to num_digits to the left of the decimal point

number	*num_digits*	ROUNDDOWN()	ROUNDUP()
1.1	0	1	2
1.678	2	1.67	1.68
1234	-2	1200	1300
-1.1	0	-1	-2
-1234	-2	-1200	-1300

Value		
3.234592654		
	Round Example	Round Statement
	3	=ROUND(A2,0)
	3.2	=ROUND(A2,1)
	3.23	=ROUND(A2,2)
	3.235	=ROUND(A2,3)
	Roundup Example	Roundup Statement
	3.24	=ROUND(A2,2)
	3.2346	=ROUND(A2,4)
	RoundDown Example	RoundDown Statement
	3.23	=ROUNDDown(A2,2)
	3.2345	=ROUNDDown(A2,4)

MROUND() is a function that rounds a number to a specified multiple:

MROUND (number, multiple)

- **number** The number you want to round
- **multiple** The multiple to which you want number rounded

- The MROUND(653, 5) rounds the number 653 to the nearest multiple of 5. The result is 655.
- MROUND(10, 3) rounds 10 to the nearest multiple of 3, resulting in 9.
- Similarly, MROUND(8, 3) results in 9 while MROUND(7, 3) results in 6 as it rounds to the nearest multiple.
- MROUND(1.5, 0.2) will round 1.5 to the nearest multiple of 0.2, that is, 1.6.
- MROUND(-1.5, -0.2) will result in -1.6.

- But MROUND(1.5, -0.2) will result in an error as the number and the multiple arguments must have the same sign.

8.3.8. CEILING.MATH() AND FLOOR.MATH()

CEILING.MATH(number[, significance, mode])

FLOOR.MATH(number[, significance, mode])

- **number** The number you want to round
- **significance** The multiple to which you want number rounded; the default value is 1 mode Determines whether Excel rounds negative numbers toward or away from zero.

CEILING is a similar function to MROUND with a difference that CEILING always rounds up the number while MROUND finds the nearest multiple. In the following table, for the same arguments, the results differ for both the functions.

	A	B	C	D	E
17	Number	Result CEILING		Result MROUND	
18	3.15	4	=CEILING(A18,1)	3	=MROUND(A18,1)
19	3.149	3.2	=CEILING(A19,0.1)	3.1	=MROUND(A19,0.1)
20					

FLOOR is similar to the CEILING function but with a difference that it always rounds number down, towards zero, to the nearest multiple of significance.

	A	B	C	D	E
17	Number	Result CEILING		Result FLOOR / MROUND	
18	3.15	4	=CEILING(A18,1)	3	=FLOOR(A18,1)
19	3.149	3.2	=CEILING(A19,0.1)	3.1	=FLOOR(A19,0.1)
20	1.58	2	=CEILING(A20,2)	0	=FLOOR(A20,2)
21				2	=MROUND(A20,2)
22	2.58	4	=CEILING(A22,2)	2	=FLOOR(A22,2)
23				2	=MROUND(A22,2)

Interesting results are of number 1.58 and 2.58. Since CEILING rounds up 1.58 to the multiple of 2, hence it returns 2 while FLOOR rounds down so returns 0 and MROUND finds the nearest multiple of 2 so returns 2. Now, incase of number 2.58, CEILING returns 4 while both FLOOR and MROUND return 2.

8.3.9. INT AND TRUNC FUNCTION

The INT() and TRUNC() functions are similar in that you can use both of them to convert a value to its integer portion

INT(number)

- **number** The number you want to round.

TRUNC(number[, num_digits])

- **number** The number you want to round.
- **num_digits** An integer that specifies the number of digits you want number rounded to, as explained here: num_digits Description

  > 0 Truncates all but num_digits decimal places

  0 Truncates all decimal places (the default)

  < 0 Converts num_digits to the left of the decimal point into zeros

For example, INT(6.75) returns 6, and TRUNC(3.6) returns 3. However, these functions have two major differences that you should keep in mind:

- For negative values, INT() returns the next number away from 0. For example, INT(-3.42) returns -4. If you just want to lop off the decimal part, you need to use TRUNC() instead.
- You can use the TRUNC() function's second argument (num_digits) to specify the number of decimal places to leave on. For example, TRUNC(123.456, 2) returns 123.45, and TRUNC(123.456, -2) returns 100.

B2 fx

	A	B	C
1		INT	TRUNC
2	2.5	2	2
3	2	2	2
4	1.5	1	1
5	1	1	1
6	0.5	0	0
7	0	0	0
8	-0.5	-1	0
9	-1	-1	-1
10	-1.5	-2	-1
11	-2	-2	-2
12	-2.5	-3	-2

8.3.10. EVEN AND ODD FUNCTIONS

EVEN() and ODD() are the last of Excel's rounding functions. They accept a single number and round it up to the nearest even or odd number.

EVEN(number)

ODD(number)

- **number** The number you want to round

To understand these quirky functions a little better, consider the following formula:

=ODD(2.6)

This formula produces the expected result: It rounds 2.6 up to the closest odd number, 3. Now consider:

=ODD(3.6)

This formula also rounds up to the nearest odd number, which in this case is 5.

In fact, ODD() always rounds up, unless you begin with a whole odd number. That means that the result of the following formula is also 5, even though 3.1 is clearly much closer to 3 than 5:

=ODD(3.1)

The EVEN() function behaves similarly. Thus, the result of the following formula is 4:

=EVEN(2.1)

The EVEN() and ODD() functions aren't useful too often. For most people, they simply represent an interesting footnote in Excel functions.

B2 =INT(A2)

	A	B	C	D	E
1		INT	TRUNC	EVEN	ODD
2	2.5	2	2	4	3
3	2	2	2	2	3
4	1.5	1	1	2	3
5	1	1	1	2	1
6	0.5	0	0	2	1
7	0	0	0	0	1
8	-0.5	-1	0	-2	-1
9	-1	-1	-1	-2	-1
10	-1.5	-2	-1	-2	-3
11	-2	-2	-2	-2	-3
12	-2.5	-3	-2	-4	-3
13					

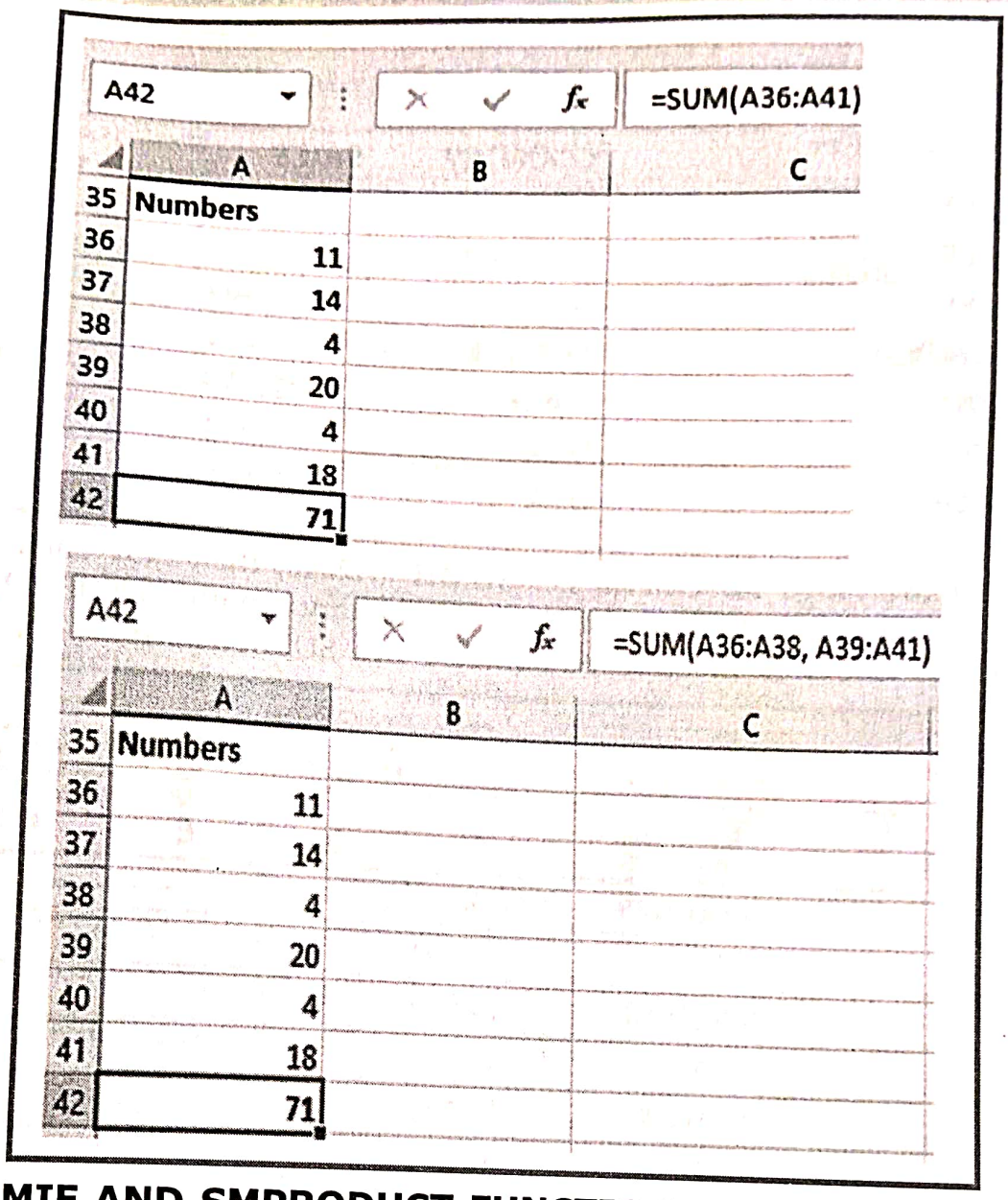
8.3.11. SUM FUNCTION

SUM(number1[, number2, ...])

- **number1, number2, ...** The values you want to add

You can enter up to 255 arguments into the SUM() function. Note, too, that you can enter the SUM() quickly by selecting the AutoSum command that appears in both the Home tab and the Formulas tab. For example, the following formula returns the sum of the values in three separate ranges:

=SUM(A2:A13, C2:C13, E2:E13)



8.3.12. SUMIF AND SMPRODUCT FUNCTION

The SUMIF() function is similar to SUM(), except that it sums only those cells in a range that meet a specified condition:

SUMIF(range, criteria[, sum_range])

- **range** The range of cells to use for the criteria.
- **criteria** The criteria, entered as text, that determines which cells to sum. Excel applies the criteria to range.
- **sum_range** The range from which the sum values are taken. Excel sums only those cells in sum_range that correspond to the cells in range and meet the criteria. If you omit sum_range, Excel uses range for the sum.

In the following figure month-wise sales data is given. If we need to sum up all the sales made in the month of January then Sumif () function can solve the problem. Below is the sample rever

	A	B	C	D
1	MONTH	COUNTRY	SALES	
2	JAN	USA	23,000	
3	FEB	ENGLAND	12,990	
4	JAN	USA	12,987	
5	MARCH	FRANCE	32,200	
6	JAN	ITALY	32,150	
7	FEB	USA	33,212	
8	JAN	USA	12,900	

First select the range for criteria i.e. SUMIF (A2:A8, then next enter the criteria i.e. JAN (as we need to sum the sales made in the January month) SUMIF(A2:A8, "JAN", and finally which values or range to be summed up) SUMIF(A2:A8, "JAN", C2:C8).

B9 =SUMIF(A2:A8, "JAN",C2:C8)

	A	B	C	D	E	F	G
1	MONTH	COUNTRY	SALES				
2	JAN	USA	23,000				
3	FEB	ENGLAND	12,990				
4	JAN	USA	12,987				
5	MARCH	FRANCE	32,200				
6	JAN	ITALY	32,150				
7	FEB	USA	33,212				
8	JAN	USA	12,900				
9	TOTAL SALES FOR JAN	81,037					

The SUMPRODUCT function provides a sophisticated way to add various products across ranges values. It doesn't just add the products of separate ranges; it produces of the value positioned in the same place in each range and then sums up those products. The syntax of the function follows:

SUMPRODUCT(range1, range2, ...)

The arguments to SUMPRODUCT must be ranges, although a range can be a single cell or value. All the range must be the same size, both rows and columns. At least 2 ranges are required and maximum up to 255 ranges are allowed.

Consider following sales data of different appliances.

	A	B	C
17	Item Name	Sold Items	Sale Price
18	TV	24500	25000
19	MOBILE	16700	22000
20	LAPTOP	2500	45000
21	AC	5500	40000
22	WASHING MACHINE	3000	24000

If want to calculate total sale then we have to calculate total sale of each item and then sum the sales of each appliance. But SUMPRODCUT make it easy as shown below:

C23 =SUMPRODUCT(B18:B22,C18:C22)

	A	B	C	D
17	Item Name	Sold Items	Sale Price	
18	TV	24500	25000	
19	MOBILE	16700	22000	
20	LAPTOP	2500	45000	
21	AC	5500	40000	
22	WASHING MACHINE	3000	24000	
23		Total Sale	1384400000	

8.3.13. SUBTOTAL

The SUBTOTAL() function is the perfect solution for all the calculations in the total row, including sums, averages, counts, and more. The trick is that the first argument of the SUBTOTAL() function is a numeric code that tells Excel what type of calculation it should perform, while the second argument is the range of cells for the entire table column, from the first row to the last.

In its default use, SUBTOTAL works just like the SUM function, except it throws out other instances of the SUBTOTAL function within the range being summed.

SUBTOTAL(*function_num, ref1, ref2, ...*)

The SUBTOTAL function takes the following arguments:

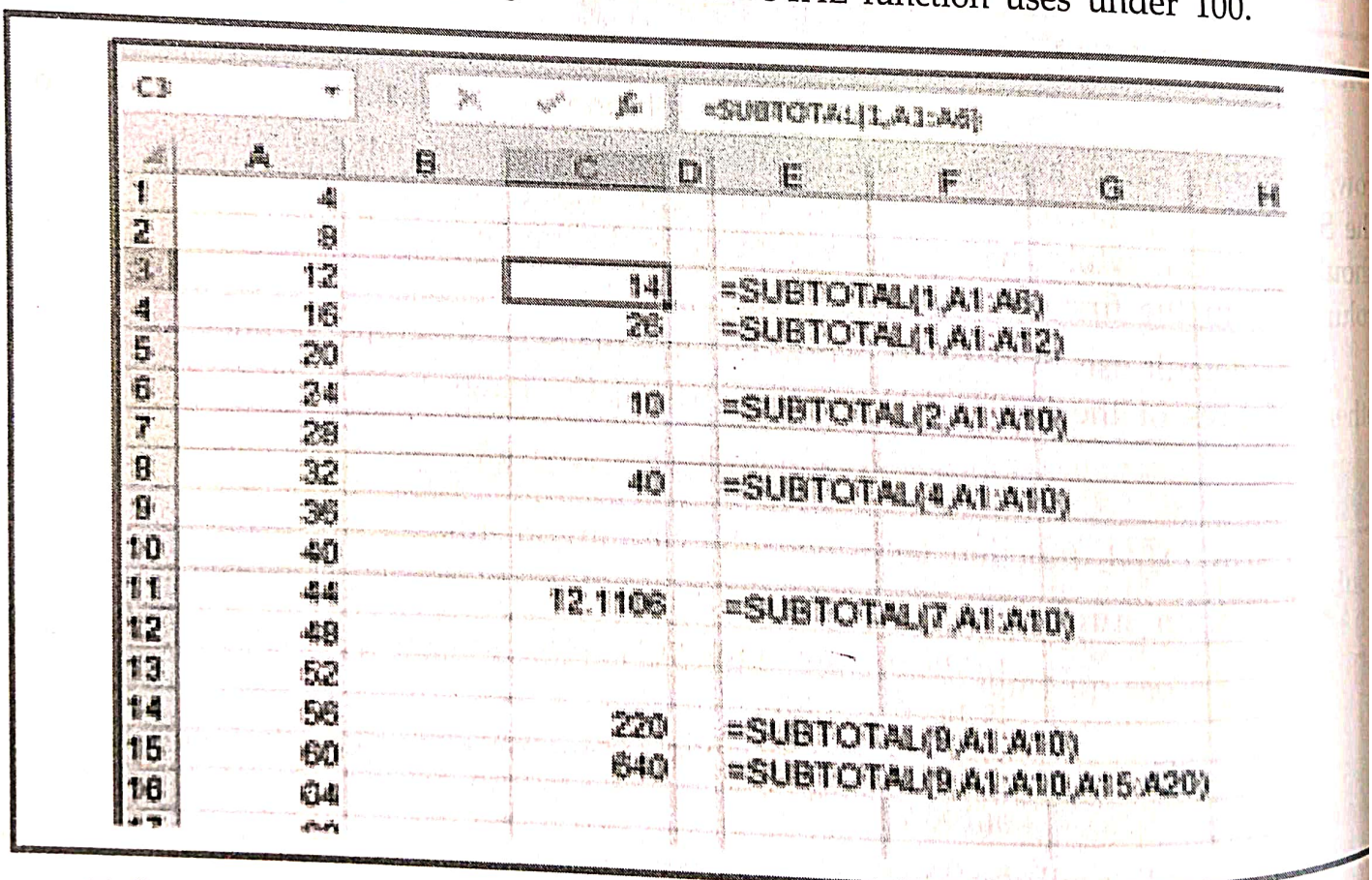
- **function_num** This is a number from 1 to 11. The most common function number is the number 9, which (for no apparent logical reason) is used to sum.
- **ref1, ref2, ...** These are up to 254 ranges or references that you want to subtotal.

Table above lists all the function codes you can use with SUBTOTAL(). Note that each calculation type actually has two function codes associated with it. Function codes above 100 ignore hidden rows (rows that you hide using the Home'! Cells'! Format'! Hide & Unhide '! Hide Rows command). These function codes are the ones the total row uses.

function_num (Includes Hidden Values)	function_num (Ignores Hidden Values)	Function
1	101	AVERAGE
2	102	COUNT
3	103	COUNTA
4	104	MAX
5	105	MIN
6	106	PRODUCT
7	107	STDEV
8	108	STDEVP
9	109	SUM
10	110	VAR
11	111	VARP

Function codes under 100 don't ignore hidden rows, but they still ignore rows you filtered out.

Consider the following figure with SUBTOTAL function uses under 100.



Following figure shows the SUBTOTAL in use with both function number system By comparing figures above and below then in the following figure row 2 is set to hidden. The values in column B are calculated using the same function Numbers as

did in above figure. In the following figure Column G shows SUBTOTAL using the function numbers start with 101. For example, cell B3 still shows the average of the numbers in the range A1:A6 equal to 15.2. The value of 8 in cell A2 is not used because it is hidden.

G3 =SUBTOTAL(101,A1:A6)

	A	B	C	D	E	F	G	H	I
1	4								
3	12		14		=SUBTOTAL(1,A1:A6)		15.2		=SUBTOTAL(101,A1:A6)
4	16		28		=SUBTOTAL(1,A1:A12)		27.63636364		=SUBTOTAL(101,A1:A12)
5	20								
6	24		10		=SUBTOTAL(2,A1:A10)		9		=SUBTOTAL(102,A1:A10)
7	28								
8	32		40		=SUBTOTAL(4,A1:A10)		40		=SUBTOTAL(104,A1:A10)
9	36								
10	40								
11	44		12.1106		=SUBTOTAL(7,A1:A10)		11.7379		=SUBTOTAL(107,A1:A10)
12	48								
13	52								
14	56		220		=SUBTOTAL(9,A1:A10)		212		=SUBTOTAL(109,A1:A10)
15	60		640		=SUBTOTAL(9,A1:A10,A15:A20)		632		=SUBTOTAL(109,A1:A10,A15:A20)
16	64								

8.4 STATISTICAL FUNCTIONS

8.4.1. COUNT, COUNTA, COUNTBLANK AND COUNTIF FUNCTIONS

All the basic counting formulas presented in this section are straightforward and relatively simple. They demonstrate the capability of the Excel counting functions to count the number of cells in a range that meet specific criteria.

To count only the numeric values in a data set (that is, to ignore text values, dates, logical values, and errors), use the COUNT() function:

COUNT(value1[, value2, ...])

- **value1, value2, ...** One or more ranges, arrays, function results, expressions, or literal values of which you want the count

In the following figure COUNT method counts the non blank numeric cell in the customer sales data.

To count not only the numeric values that appear in a data set but also the text values, dates, logical values, and errors, use the COUNTA() function:

COUNTA(value1[, value2, ...])

B9 =COUNT(B3:B8)

	A	B
1	Customers Sales Data	
2	Sales Man	Sales
3	Raman	50000
4	Amrit	65000
5	Pawan	75000
6	Guri	
7	Jasmine	95000
8	Harman	120000
9		5

B9 =COUNT(B3:B8)

	A	B
1	Customers Sales Data	
2	Sales Man	Sales
3	Raman	50000
4	Amrit	65000
5	Pawan	No Sale
6	Guri	
7	Jasmine	95000
8	Harman	120000
9		4

- **value1, value2, ...** One or more ranges, arrays, function results, expressions, or literal values of which you want the count.

In the following figure COUNTA function counts numeric and non numeric cells both.

B9 =COUNTA(B3:B8)

	A	B
1	Customers Sales Data	
2	Sales Man	Sales
3	Raman	50000
4	Amrit	65000
5	Pawan	No Sale
6	Guri	TRUE
7	Jasmine	95000
8	Harman	120000
9		6

The following formula returns the number of blank (empty) cells in a range named Data:

COUNTBLANK(value1[, value2, ...])

- **value1, value2, ...** One or more ranges, arrays, function results, expressions, or literal values of which you want the count

This function works only with a contiguous range of cells. If Data is defined as a noncontiguous range, the function returns a #VALUE! error. The COUNTBLANK function also counts cells containing a formula that returns an empty string.

In the following figure COUNT method counts the blank cells in the customer sales data.

B9 =COUNTBLANK(B3:B8)

	A	B	C
1	Customers Sales Data		
2	Sales Man	Sales	
3	Raman	50000	
4	Amrit	65000	
5	Pawan	75000	
6	Guri		
7	Jasmine	95000	
8	Harman	120000	
9		1	

The COUNTIF function, which is useful for single-criterion counting formulas, takes two arguments:

COUNTIF(*range, criteria*)

- **range:** The range that contains the values that determine whether to include a particular cell in the count
- **criteria:** The logical criteria that determine whether to include a particular cell in the count

Following table shows the examples of COUNTIF examples where 'Data' is range of cells containing values.

=COUNTIF (Data, 12)	Returns the number of cells containing the value 12
=COUNTIF (Data, "<0")	Returns the number of cells containing a negative value
=COUNTIF (Data, "<>0")	Returns the number of cells not equal to 0
=COUNTIF (Data, ">5")	Returns the number of cells greater than 5
=COUNTIF (Data, A1)	Returns the number of cells equal to the contents of cell A1

Following figure count the sales which are more than 75000 in a month.

B9 f_x =COUNTIF(B3:B8, "> 75000")

	A	B	C
1	Customers Sales Data		
2	Sales Man	Sales	
3	Raman	50000	
4	Amrit	65000	
5	Pawan	70000	
6	Guri	85000	
7	Jasmine	95000	
8	Harman	120000	
9		3	

In many cases, your counting formula will need to count cells only if two or more criteria are met. These criteria can be based on the cells that are being counted or on a range of corresponding cells.

The COUNTIFS() function counts the number of cells in one or more ranges that meet one or more criteria:

COUNTIFS(range1, criteria1[, range2, criteria2, ...])

- **range1** The first range of cells to use for the count.
- **criteria1** The first criteria, entered as text, that determines which cells to count. Excel applies the criteria to range1.
- **range2** The second range of cells to use for the count.
- **criteria2** The second criteria, entered as text, that determines which cells to count. Excel applies the criteria to range2.

You can enter up to 127 range/criteria pairs, and Excel only counts those cells that match all the criteria.

An *And* criterion counts cells if all specified conditions are met. A common example is a formula that counts the number of values that fall within a numerical range. For

B9 f_x =COUNTIFS(B3:B8, "> 70000", B3:B8, "< 960000")

	A	B	C	D	E
1	Customers Sales Data				
2	Sales Man	Sales			
3	Raman	50000			
4	Amrit	65000			
5	Pawan	70000			
6	Guri	85000			
7	Jasmine	95000			
8	Harman	120000			
9		3			

example, you may want to count cells that contain a value that's greater than 100 and less than or equal to 200. For this example, the COUNTIFS function will do the job.

```
=COUNTIFS(B3:B8, "> 70000", B3:B8, "< 96000")
```

An *Or* criterion counts cells if any of the multiple conditions is met. To count cells by using an Or criterion, you can sometimes use multiple COUNTIF functions. The following formula, for example, counts the number of sales made in January or February:

```
=COUNTIF(B3:B8, "70000")+COUNTIF(B3:B8, "95000")
```

B9 fx =COUNTIF(B3:B8, "70000")+ COUNTIF(B3:B8, "95000")

	A	B	C	D	E
1	Customers Sales Data				
2	Sales Man	Sales			
3	Raman	50000			
4	Amrit	65000			
5	Pawan	70000			
6	Guri	85000			
7	Jasmine	95000			
8	Harman	120000			
9		2			

8.4.2. MIN, MAX, SMALL AND LARGE FUNCTIONS

If you have a large data set and want to find the smallest or largest value in a column, rather than sort the data set, you can use a function to find the value. To find the smallest numeric value, you use MIN. To find the largest numeric value, you use MAX.

The MIN function returns the smallest number in a set of values.

```
MIN(number1, number2, ...)
```

- **number1, number2:**. are one to 255 numbers for which you want to find the minimum value. You can specify arguments that are numbers, empty cells, logical values, or text representations of numbers. Arguments that are error values or text that cannot be translated into numbers cause errors.

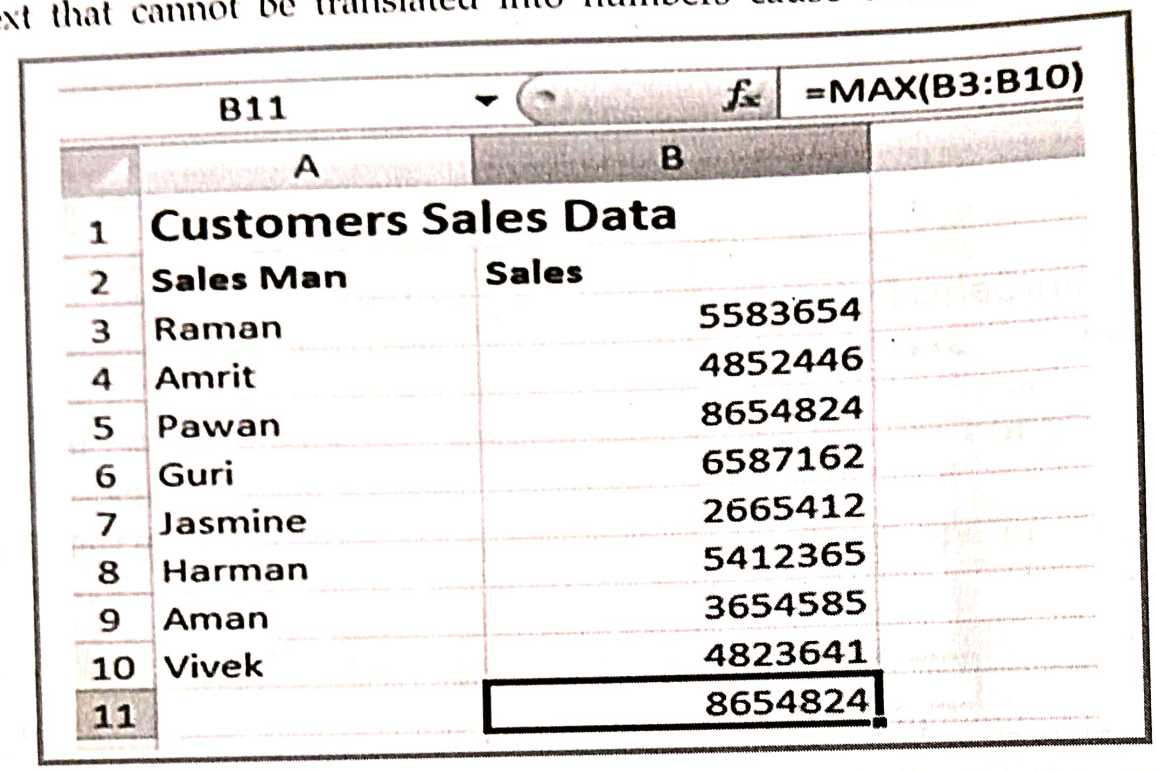
B11 fx =MIN(B3:B10)

	A	B
1	Customers Sales Data	
2	Sales Man	Sales
3	Raman	5583654
4	Amrit	4852446
5	Pawan	8654824
6	Guri	6587162
7	Jasmine	2665412
8	Harman	5412365
9	Aman	3654585
10	Vivek	4823641
11		2665412

The MAX function returns the largest value in a set of values.

MAX(*number1, number2, ...*)

- **number1, number2:**. are one to 255 numbers for which you want to find the minimum value. You can specify arguments that are numbers, empty cells, logical values, or text representations of numbers. Arguments that are error values or text that cannot be translated into numbers cause errors.



The MIN function finds the smallest value in a data set. The SMALL function can find the *k*th smallest value. This can be great for finding not just the smallest value but the second-smallest, third-smallest, and so on. The SMALL function returns the *k*th smallest value in a data set.

SMALL(*array, k*)

- **array**: is an array or a range of numeric data for which you want to determine the *k*th smallest value. If *array* is empty, SMALL returns a #NUM! error.
- **k:** is the position (from the smallest) in the array or range of data to return. If *k* is less than or equal to 0 or if *k* exceeds the number of data points, SMALL returns a #NUM! error.

To find the 3rd smallest sale in the customer data use SMALL function as given in figure.

The MAX function discussed in the preceding section finds the single largest value in a list. The LARGE function returns the *k*th largest value in a data set. You can use this function to select a value based on its relative standing. The LARGE function can find the first, second, third, and so on largest values in a list.

LARGE(*array, k*)

B12 f_x =SMALL(B3:B11,3)

	A	B	C
1	Customers Sales Data		
2	Sales Man	Sales	
3	Raman	5583654	
4	Amrit	4852446	
5	Pawan	8654824	
6	Guri	6587162	
7	Jasmine	2665412	
8	Harman	5412365	
9	Aman	3654585	
10	Vivek	4823641	
11	Sahib	8654824	
12		4823641	

This function takes the following arguments:

- *array*: This is the array or range of data for which you want to determine the *k*th largest value. If *array* is empty, LARGE returns a #NUM! error.
- *k*: This is the position (from the largest) in the array or cell range of data to return. If *k* is less than or equal to 0, or if *k* is greater than the number of data points, LARGE returns a #NUM! error.

To find the 3rd largest sale in the customer data use LARGE function as given in figure.

B12 f_x =LARGE(B3:B11,3)

	A	B	C
1	Customers Sales Data		
2	Sales Man	Sales	
3	Raman	5583654	
4	Amrit	4852446	
5	Pawan	8654824	
6	Guri	6587162	
7	Jasmine	2665412	
8	Harman	5412365	
9	Aman	3654585	
10	Vivek	4823641	
11	Sahib	8654824	
12		6587162	
13			

8.4.3. MEAN FUNCTIONS

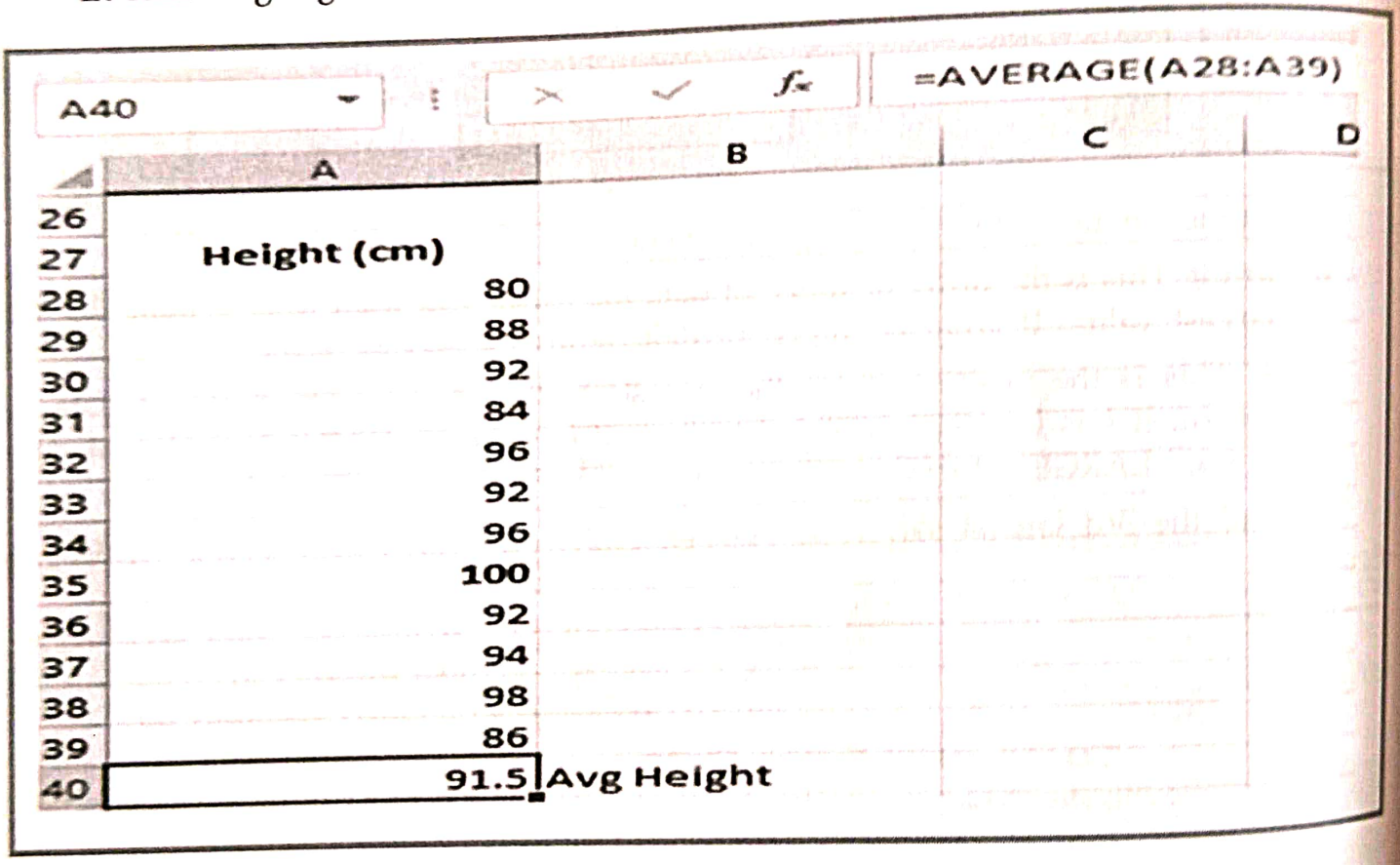
The mean of a dataset is the mathematical average. It is calculated by adding all the values in the range and dividing by the number of values in the set. To calculate a mean in Excel, you use the AVERAGE function.

The AVERAGE function returns the average (that is, arithmetic mean) of the arguments.

AVERAGE(number1, number2, ...)

- **number1, number2**, ... are 1 to 255 numeric arguments for which you want the average. The arguments must be either numbers or names, arrays, or references that contain numbers. If an array or a reference argument contains text, logical values, or empty cells, those values are ignored; however, cells containing the value 0 are included.

In following Figure calculate the average height of boys in a college.



In some data sets, one value might be more important than another. You can do this by calculating the weighted mean. This is an arithmetic mean in which each value is weighted according to its importance in the data set. SUMPRODUCT() works by multiplying the corresponding items in each range or array (that is, the first items are multiplied, then the second items are multiplied, and so on) and then summing those products.

.SUMPRODUCT(array1[, array2, ...])

- **array1, array2**, ... The ranges or arrays you want to work with, each of which must have the same number of elements

Following figure shows a worksheet that lists the profit margin (B2:B6) and sales (C2:C6) for four cities. Cell C8 uses the following formula to calculate the weighted mean profit margin:

C8 =SUMPRODUCT(B2:B6,C2:C6)/SUM(C2:C6)

	A	B	C	D	E	F	G
1	Cities	Profit %	Sales				
2	Ludhiana	4.50%	10000000				
3	Jalandhar	7.50%	5000000				
4	Amritsar	9.50%	2000000				
5	Mohali	5.50%	4000000				
6	Patiala	8.50%	200000				
7							
8	Weighted Mean Pro		5.91%				

The GEOMEAN function returns the geometric mean of an array or a range of positive data.

For example, you can use GEOMEAN to calculate average growth rate, given compound interest with variable rates.

GEOMEAN(number1, number2, ...)

- **number1, number2**, ...: are 1 to 255 arguments for which you want to calculate the mean. You can also use a single array or a reference to an array instead of arguments separated by commas.

The arguments must be either numbers or names, arrays, or references that contain numbers. If an array or a reference argument contains text, logical values, or empty cells, those values are ignored; however, cells that contain the value 0 are included. If any data point is less than or equal to 0, GEOMEAN returns a #NUM! error.

Use the GEOMEAN when you want to calculate the average growth rate of compound interest, assuming variable interest rates. Following figure shows an example.

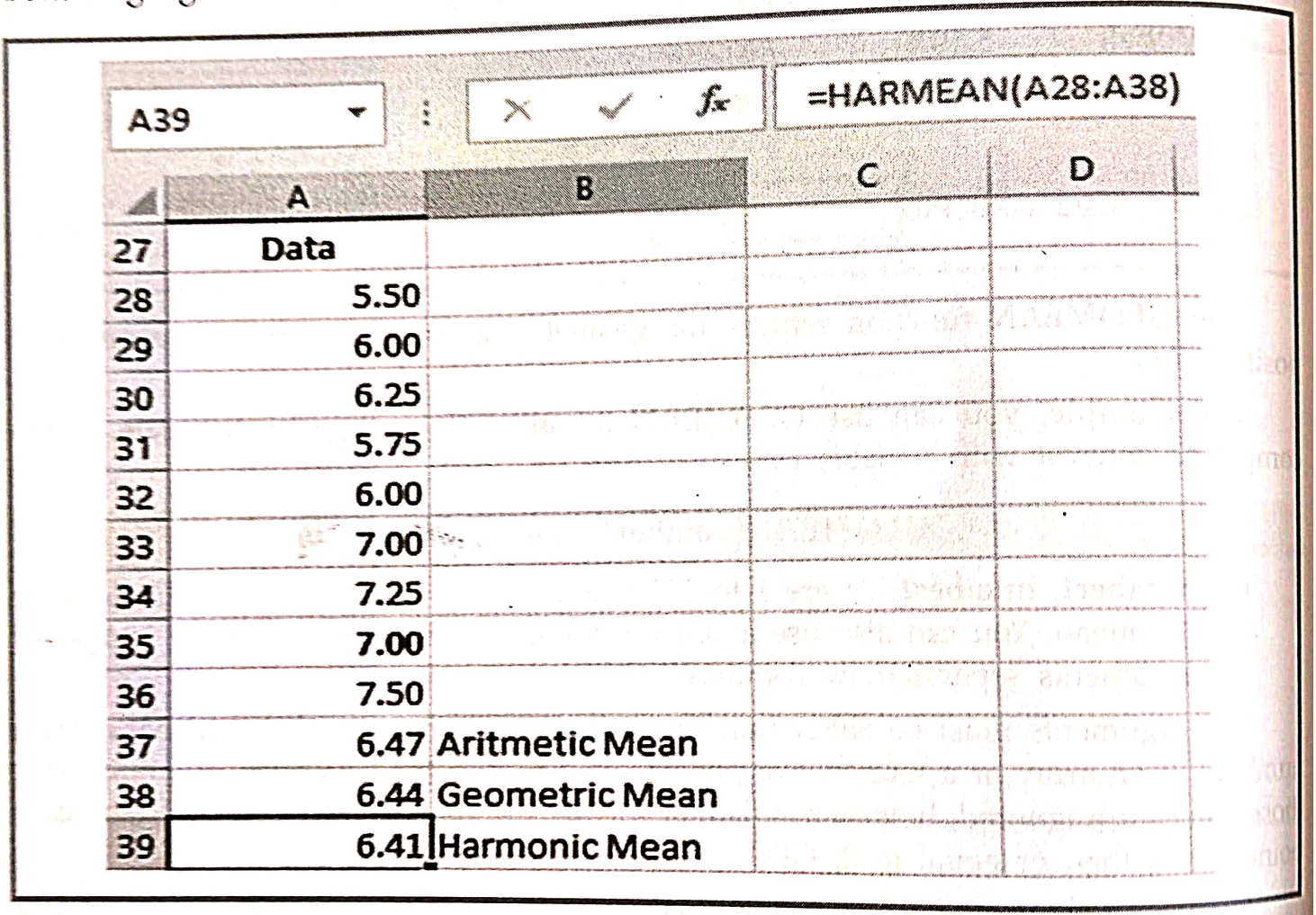
Interest_Rate	Mean	
5.50	7.07	=AVERAGE(A2:A8)
6.50		
6.75		
7.00	GeoMean	
7.50	7.02	=GEOMEAN(A2:A8)
8.00		
8.25		

The harmonic mean is the reciprocal of the arithmetic mean of reciprocals. HARMEAN returns the harmonic mean of a data set.

HARMEAN(number1, number2, ...).

- number1, number2, ...are 1 to 255 arguments for which you want to calculate the mean. You can also use a single array or a reference to an array instead of arguments separated by commas.

The harmonic mean is always less than the geometric mean and arithmetic mean. Following figure shows the concept.



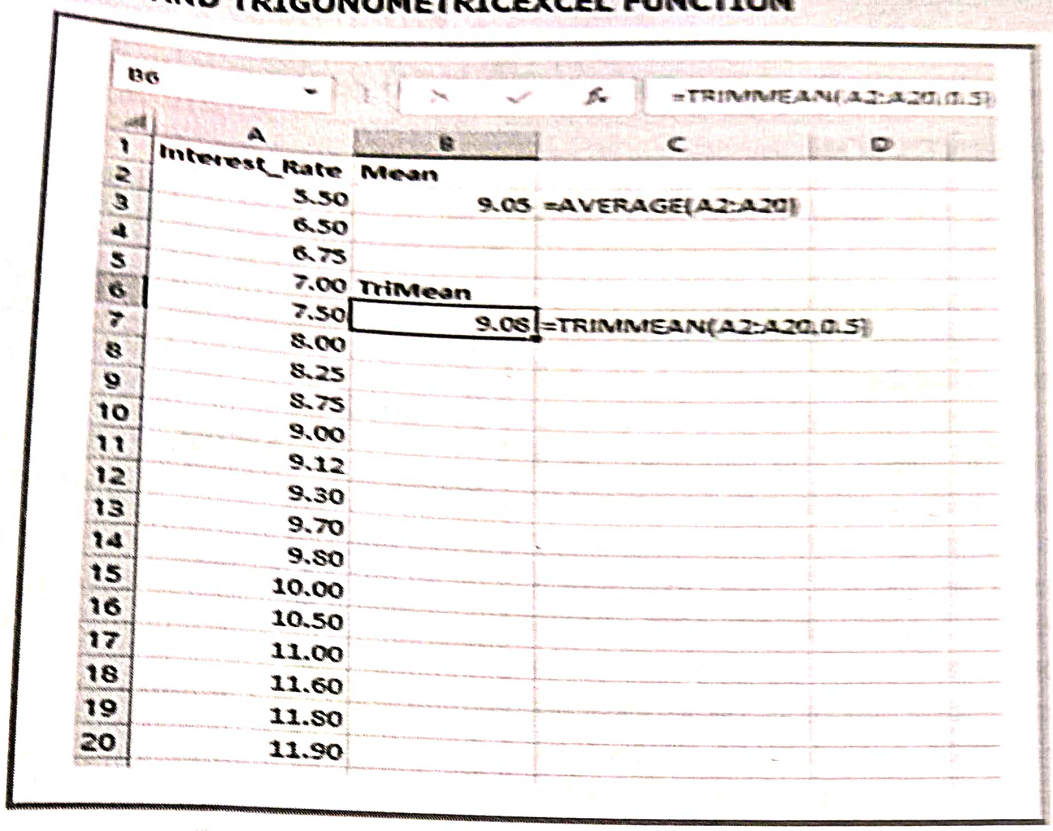
	A	B	C	D
27	Data			
28	5.50			
29	6.00			
30	6.25			
31	5.75			
32	6.00			
33	7.00			
34	7.25			
35	7.00			
36	7.50			
37	6.47	Aritmetic Mean		
38	6.44	Geometric Mean		
39	6.41	Harmonic Mean		

The TRIMMEAN function is used when you want to exclude data points from the top and bottom of the data set, for example when the data set contains unusually high or low values which skew the mean. TRIMMEAN returns the mean of the interior of a data set.

TRIMMEAN(array, percent)

- array: Is the data set, which can be cell references, arrays, or range names.
- percent: Is the fractional number of data points to exclude from the calculation.

In Figure, the percent argument is 0.50% and the data set contains 20 values; 0.50% of 20 is 10. The TRIMMEAN function rounds down to the nearest multiple of 2. In this example, it will trim four values from the top and four from the bottom.



	A	B	C	D
1	Interest_Rate	Mean		
2				
3	5.50	9.05	=AVERAGE(A2:A20)	
4	6.50			
5	6.75			
6	7.00	TriMean		
7	7.50	9.08	=TRIMMEAN(A2:A20,0.5)	
8	8.00			
9	8.25			
10	8.75			
11	9.00			
12	9.12			
13	9.30			
14	9.70			
15	9.80			
16	10.00			
17	10.50			
18	11.00			
19	11.60			
20	11.80			
	11.90			

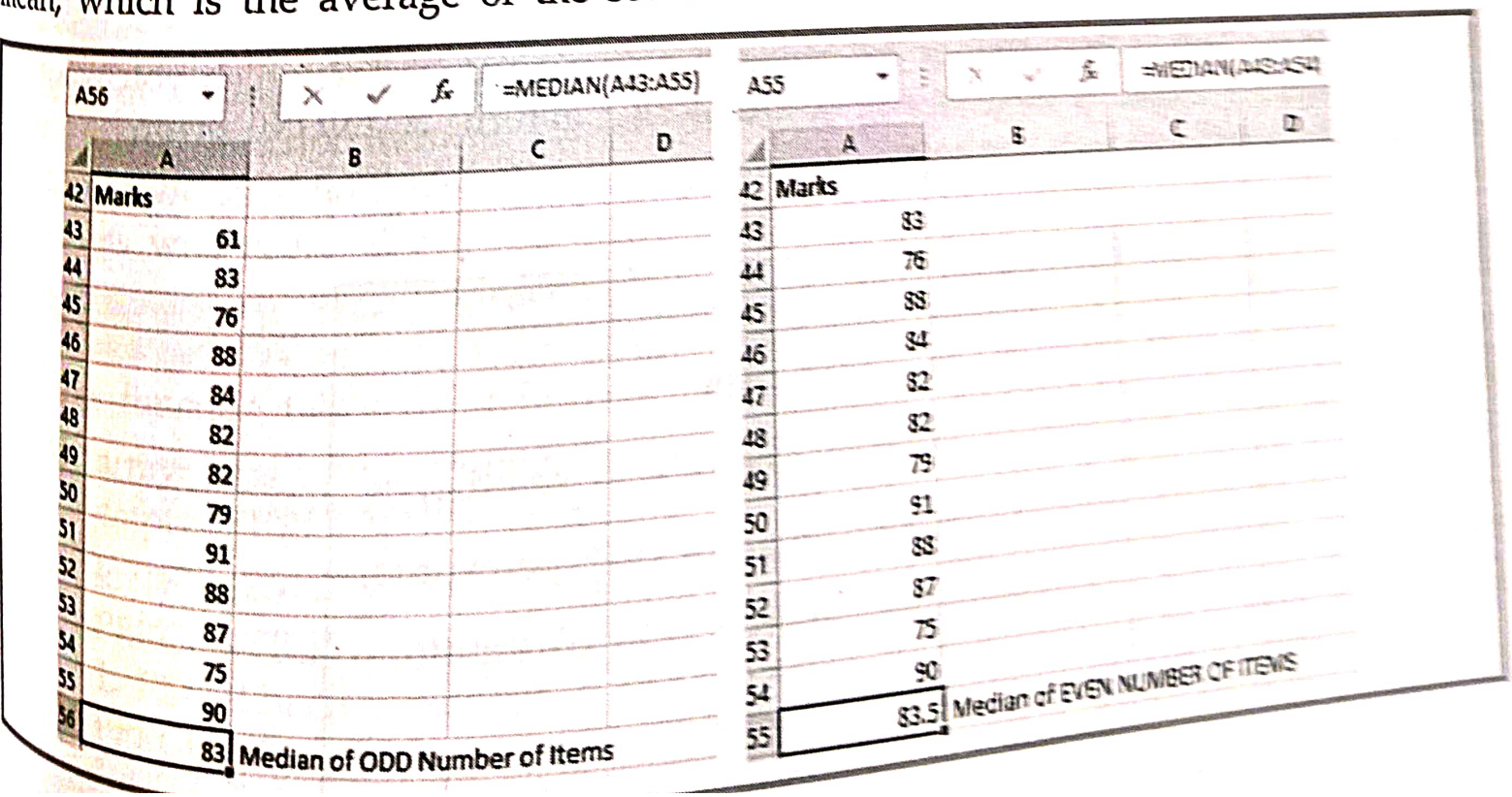
8.4.4. MEDIAN FUNCTION

MEDIAN returns the median of the given numbers.

MEDIAN(number1, number2, ...)

- **number1, number2, ...** are 1 to 255 numbers for which you want the median. The arguments should be either numbers or names, arrays, or references that contain numbers.

The median is the number in the middle of the set of numbers, as opposed to the mean, which is the average of the set of numbers. When the data set contains an odd



	A	B	C	D
42	Marks			
43	61			
44	83			
45	76			
46	88			
47	84			
48	82			
49	82			
50	79			
51	91			
52	88			
53	87			
54	75			
55	90			
56	83	Median of ODD Number of Items		

	A	B	C	D
42	Marks			
43	83			
44	76			
45	88			
46	84			
47	82			
48	82			
49	79			
50	91			
51	88			
52	87			
53	75			
54	90			
55	83.5	Median of EVEN NUMBER OF ITEMS		

number of values, the value in the middle is returned. When the data set contains an even number of values, the middle two values are averaged to determine the median.

8.4.5. MODE FUNCTION

MODE returns the most frequently occurring, or repetitive, value in an array or range of data.

MODE(number1, number2, ...)

number1, number2, ...: are 1 to 255 arguments for which you want to calculate the mean. You can also use a single array or a reference to an array instead of arguments separated by commas Use MODE function when you need to know which value occurs most frequently in a data set. If each value is unique, there is no mode and the error #NA is returned. If two or more values occur with the same frequency, Excel returns the item with the smallest value regardless of the order in which it encounters the values in the data set. Following figure shows an example of MODE.

Set 1		
4	MODE	
2	9	=MODE(A2:A9)
9		
7		
6		
9		
3		
9		

8.4.6. QUARTILE FUNCTIONS

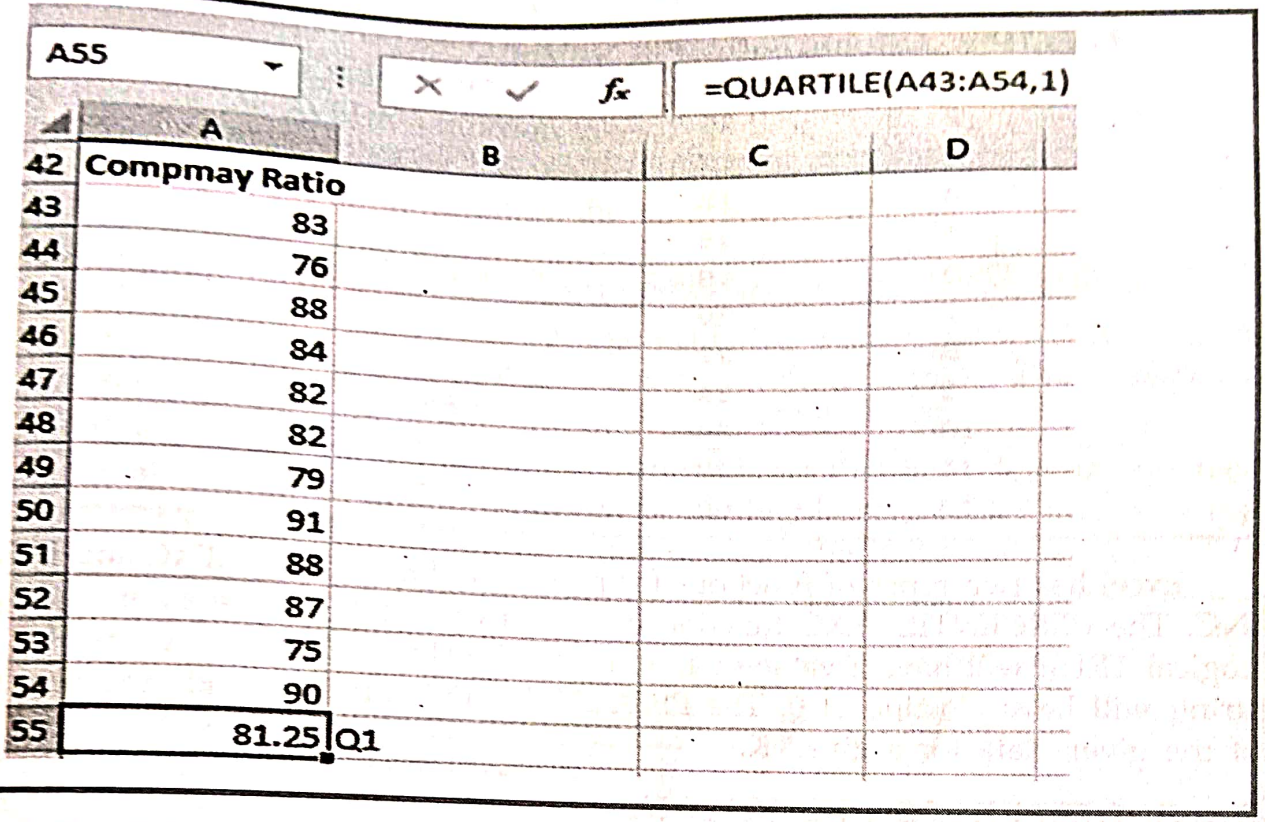
The QUARTILE function is used frequently to divide data into four equal parts and return the average of the quartile you specify.. Quartiles are often used in sales and survey data to divide populations into groups.

QUARTILE(array, quart)

- *array*: This is the array or cell range of numeric values for which you want the quartile value. If *array* is empty, QUARTILE returns a #NUM! error.
- *quart* This indicates which value to return. Following table shows the quart values.

Quart Value	Result
0	Minimum value
1	First quartile (25th percentile)
2	Second quartile (50th percentile)
3	Third quartile (75th percentile)
4	Maximum value

In following figure, the QUARTILE function is being used to locate the top twenty-fifth percentile of price-earnings ratios for 12 companies.



	A	B	C	D
42	Compmay Ratio			
43	83			
44	76			
45	88			
46	84			
47	82			
48	82			
49	79			
50	91			
51	88			
52	87			
53	75			
54	90			
55	81.25	Q1		

In Excel, the QUARTILE function is classified into two types, QUARTILE.EXC and QUARTILE.INC. The QUARTILE.EXC function excludes the zeroth quartile and fourth quartile of the given data for computing the quartiles, whereas QUARTILE.INC includes the zeroth quartile and fourth quartile along with first quartile, second quartile, and third quartile of the given data.

8.4.7. PERCENTILE FUNCTION

The QUARTILE function is fine if you are trying to find every record that is in the top 25% of a range. Sometimes, however, you need to find some other percentile. The PERCENTILE function returns the kth percentile of values in a range. You can use this function to establish a threshold of acceptance. For example, you can decide to examine candidates who score above the 90th percentile.

PERCENTILE(array, k)

This function takes the following arguments:

- **array:** This is the array or range of data that defines relative standing. If array is empty, PERCENTILE returns a #NUM! error.
- **k:** This is the percentile value in the range 0...1, inclusive. If k is non numeric, PERCENTILE returns a #VALUE! error. If k is less than 0 or if k is greater than 1, PERCENTILE returns a #NUM! error. If k is not a multiple of 1 / (n – 1), PERCENTILE interpolates to determine the value at the kth percentile.

In the following example 12 companies has Earning ratio. If we have to find the percentile range of 70% i.e. 0.7 companies instead of 100% then we use the k value.

Company	Earning Ratio	Percentile	
1	7	28.3	=PERCENTILE(B2:B13,0.7)
2	16		
3	17		
4	18		
5	19		
6	19		
7	20		
8	22		
9	31		
10	34		
11	38		
12	60		

Excel has two types of functions for percentile, PERCENTILE.EXC and PERCENTILE.INC. The PERCENTILE.EXC function determines percentile value by excluding 0 and 1. Logical TRUE will have a value of 1, and logical FALSE will have a value of 0. Any other string will have a value of 0. The PERCENTILE.INC function finds the percentile value of the given data for a given K.

8.4.8. DISPERSION FUNCTIONS

Dispersion in statistics is a way to describe how spread out or scattered the data is around an average value. It helps to understand if the data points are close together or far apart. Dispersion shows the variability or consistency in a set of data. There are different measures of dispersion like range, variance, and standard deviation.

1. **Range:** Excel Formula: = MAX-MIN. The difference between the maximum value and the minimum value in the data set.
2. **Interquartile Range:** It is the difference between the first and third quartile; this measure included only the middle 50% of the data and, therefore, is not influenced by extreme values. You can use QUARTILE function to calculate Q1 and Q3 then find interquartile range using (Q3 – Q1)/2.
3. **Mean Deviation:** The Excel AVEDEV function returns the average of the absolute value of deviations from the mean for a given set of data. Mean deviation is a measure of variability.
4. **Standard Deviation:** Square root of the variance in Excel, for sample = STDEV.S(data, range). A high standard deviation means more variability in the data. For population = STEDV.P(data, range).
5. **Variance:** The 'average' of the squared deviation from the mean. Excel function for sample = VAR.S (data, range). For population = VAR.P(data, range). The formula for population divides by N, whereas the formula for sample divides by N-1.

The AVEDEV function is one of several functions used to measure the dispersion, or spread, in a set of data. AVEDEV returns the average of the absolute deviations of data points from their mean. AVEDEV is influenced by the unit of measurement in the input data.

The equation for the AVEDEV function takes 1 divided by the number of values in a data set (n), then multiplies that by the result of the sum of the absolute value of each data point (x1, x2, ..., xn) less the mean of the data set.

AVEDEV(number1, number2, ...)

- **number1, number2, ...**: are 1 to 255 arguments for which you want to calculate the mean. You can also use a single array or a reference to an array instead of arguments separated by commas.

Following figure shows the AVERAGE calculation shows that both departments have an average age of 40. But the AVEDEV calculation clearly highlights a wider spread in ages for Department B, the average deviation is 17 points from the mean (average)

B49 | =AVEDEV(B43:B47)

	A	B	C	D
42	Department A	Department B		
43	39	69		
44	42	32		
45	35	18		
46	41	52		
47	36	28		
48	43	40	Avg of Dept B	
49	42	17	Mean Dev of Dept B	
50	40	Avg of Dept A		
51	3	Mean Dev of Dept A		

The standard deviation is a measure of how widely dispersed values are from the arithmetic mean (the average) value. In Excel, the standard deviation of a set of observations of a sample and that of a set of observations of a population can be obtained using the following functions, respectively. The STDEV.S function estimates standard deviation based on a sample. it containing data of sample excluding logical and text.

STDEV.S(number1, [number2], ...)

- **number1, number2, ...**: are 1 to 255 arguments for which you want to calculate the mean. You can also use a single array or a reference to an array instead of arguments separated by commas.

Following figure shows the life cycle of two products. It clearly shows that life of product 1 is more variable than product 2.

A71 =STDEV.S(A56:A70)

	A	B	C	D
55	Life of Product	Life of Product 2		
56	68	77		
57	70	80		
58	71	81		
59	72	81		
60	74	82		
61	74	83		
62	78	83		
63	82	84		
64	83	84		
65	88	85		
66	90	88		
67	92	89		
68	93	90		
69	96	92		
70	97	95		
71	10.2	4.92		

The STDEV.P function calculates standard deviation based on the entire population, given as arguments. STDEV.P assumes that its arguments are the entire population. If your data represents a sample of the population, you can compute the standard deviation by using STDEV.S. For large sample sizes, STDEV.S and STDEV.P return approximately equal values. The standard deviation is calculated using the "biased" or "n" method. It containing data of population excluding logical and text.

STDEV.P(number1, [number2], ...)

- **number1, number2, ...**: are 1 to 255 arguments for which you want to calculate the mean. You can also use a single array or a reference to an array instead of arguments separated by commas.

A71 =STDEV.P(A56:A70)

	A	B	C	D
55	Life of Product	Life of Product 2		
56	68	77		
57	70	80		
58	71	81		
59	72	81		
60	74	82		
61	74	83		
62	78	83		
63	82	84		
64	83	84		
65	88	85		
66	90	88		
67	92	89		
68	93	90		
69	96	92		
70	97	95		
71	9.8	4.75		

Using the same two products life data to calculate the standard deviation then see difference in the results.

Another types of standard deviation function are:

STDEVA estimates standard deviation based on a sample that may be comprised of numbers, text, or logical values. The difference is that STDEVA does not ignore text or logical values (TRUE and FALSE). Text and FALSE are evaluated as zero (0) and TRUE is evaluated as one (1).

STDEVA(value1, value2, ...)

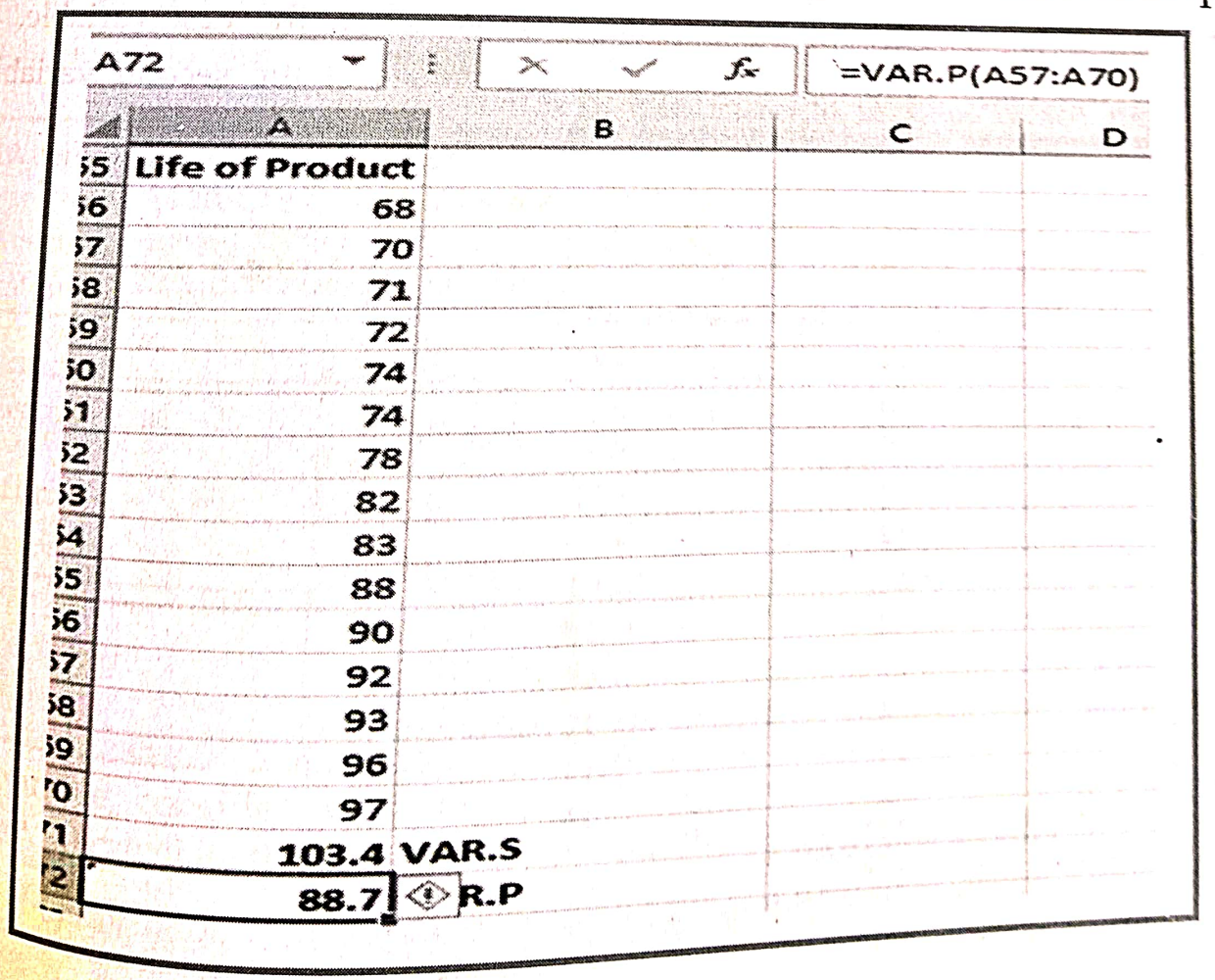
STDEVPA calculates standard deviation based on the entire population given as arguments.

The STDEVPA function differs from the STDEVP function in that it will also evaluate logical values such as TRUE and FALSE. TRUE evaluates to one (1); FALSE evaluates to zero (0). Arguments that contain text also evaluate to zero.

STDEVPA(value1, value2, ...)

Variance is a measure of how spread out set of data is in relation to the mean, Variance is calculated by summing the squared deviations from the mean. Variance is difficult to interpret and to apply to a set of data. But variance is important because it is used both as a concept and as a practical measure of variability in manly statistics formula and technique.

Variance is commonly used in finance, and many statistical calculations, but a key ssue that must be resolved is whether the user is looking for the variance of the population



or the variance of a representative sample of the larger data set. That is the difference between VAR.P—population—and VAR.S—sample.

VAR.P(number1, number2, ...)
VAR.S(number1, number2, ...)

- **number1, number2, ...**: are 1 to 255 arguments for which you want to calculate the mean. You can also use a single array or a reference to an array instead of arguments separated by commas.

Following figure calculates the variance of product life as sample and as a population. See the difference in the result.

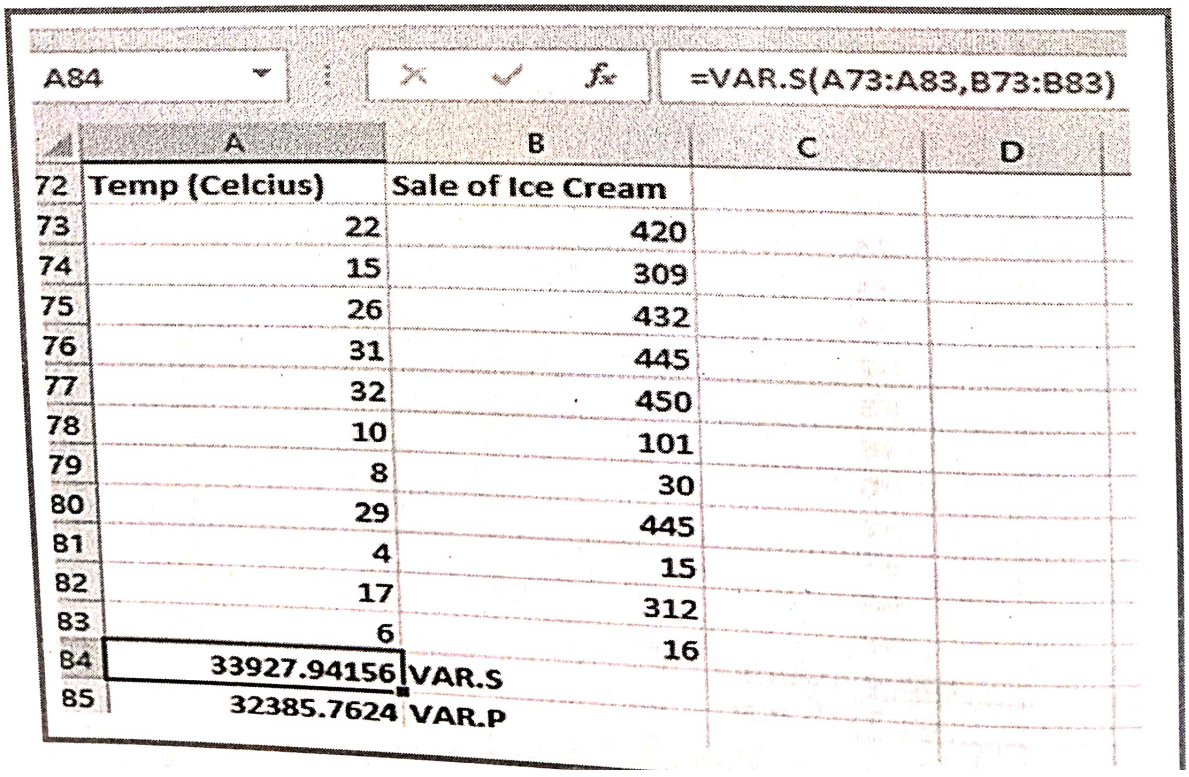
8.4.9. COVARIANCE AND CORRELATION FUNCTIONS

Correlation and covariance are two statistical concepts that are used to determine the relationship of two random variables. Covariance measures the total variation of two random variables from their expected values, help to gauge the direction of the relationship, that is, whether the variables tend to move in tandem (positive covariance) or show an inverse (negative covariance) relationship.

Excel has two functions, COVARIANCE.P and COVARIANCE.S, for covariance of the whole population and a sample of the population, similar to variance and standard deviation functions.

COVARIANCE.S(array1, array2)
COVARIANCE.P(array1, array2)

- **array1** The range or array that contains the values of the first variable.
- **array2** The range or array that contains the values of the second variable.



Interpretation of Result:

- If the covariance value is a positive number, it means the two variables have a direct relationship.
- If the covariance value is a negative number, it means the two variables have an inverse relationship.
- If the covariance value is 0 (or close to it), it means the two variables are not related.

Therefore, you can say that temperatures and sales of ice cream have a direct relationship with each other

Correlation measures the strength of the relationship between variables, how change in one impacts the other. A correlation of +1 indicates that random variables have a direct and strong relationship, -1 indicates that there is a strong inverse relationship and an increase in one variable will lead to an equal and opposite decrease in the other variable and 0 indicates that the two numbers are independent. The **CORREL()** function calculates the *correlation coefficient*.

CORREL(array1, array2)

- **array1** The range or array that contains the values of the first variable.
- **array2** The range or array that contains the values of the second variable.

The coefficient is a number between –1 and 1 that has the following properties:

Correlation Coefficient	Interpretation
1	The two variables have a direct relationship that is perfectly and positively correlated. For example, a 10% increase in temperature a 10% increase in sales of ice cream.
Between 0 and 1	The two variables have a direct relationship. For example, an increase in temperature leads to an increase in sales of ice cream. The higher the number, the stronger the direct relationship.
0	There is no relationship between the variables.
Between 0 and –1	The two variables have an inverse relationship. For example, an increase in temperature leads to a decrease in sales of ice cream. The lower the number, the stronger the inverse relationship.
–1	The variables have an inverse relationship that is perfectly and negatively correlated. For example, a 10% increase in temperature leads to a 10% decrease in sales of ice cream.

For example, consider again the advertising and sales numbers shown earlier in figure above. To calculate the correlation between these variables, use the following formula:

=CORREL(A73:A83, B73:B83).

A84 =CORREL(A73:A83,B73:B83)

	A	B	C	D
72	Temp (Celcius)	Sale of Ice Cream		
73	22	420		
74	15	309		
75	26	432		
76	31	445		
77	32	450		
78	10	101		
79	8	30		
80	29	445		
81	4	15		
82	17	312		
83	6	16		
84	0.95			

The result, as you can see in cell A84 in Figure is 0.95, which tells you not only that temperature and sales of ice cream have a direct relationship, but also that this relationship is relatively strong.

8.4.10. REGRESSION FUNCTIONS

Regression analysis allows you to predict the future, based on past events. Say that you have observed total sales for the past several years. Regression analysis finds a line

A84 =FORECAST(600,A73:A83,B73:B83)

	A	B	C	D	E
71					
72	Temp (Celcius)(Y)	Sale of Ice Cream (X)			
73	22	420			
74	15	309			
75	26	432			
76	31	445			
77	32	450			
78	10	101			
79	8	30			
80	29	445			
81	4	15			
82	17	312			
83	6	16			
84	35	600			

that best fits the past data points. You can then use the description of that line to predict results for the future data points.

Use the **FORECAST** function when you want to predict a value based on the trend of the existing or known data.

FORECAST(x, known_y's, known_x's)

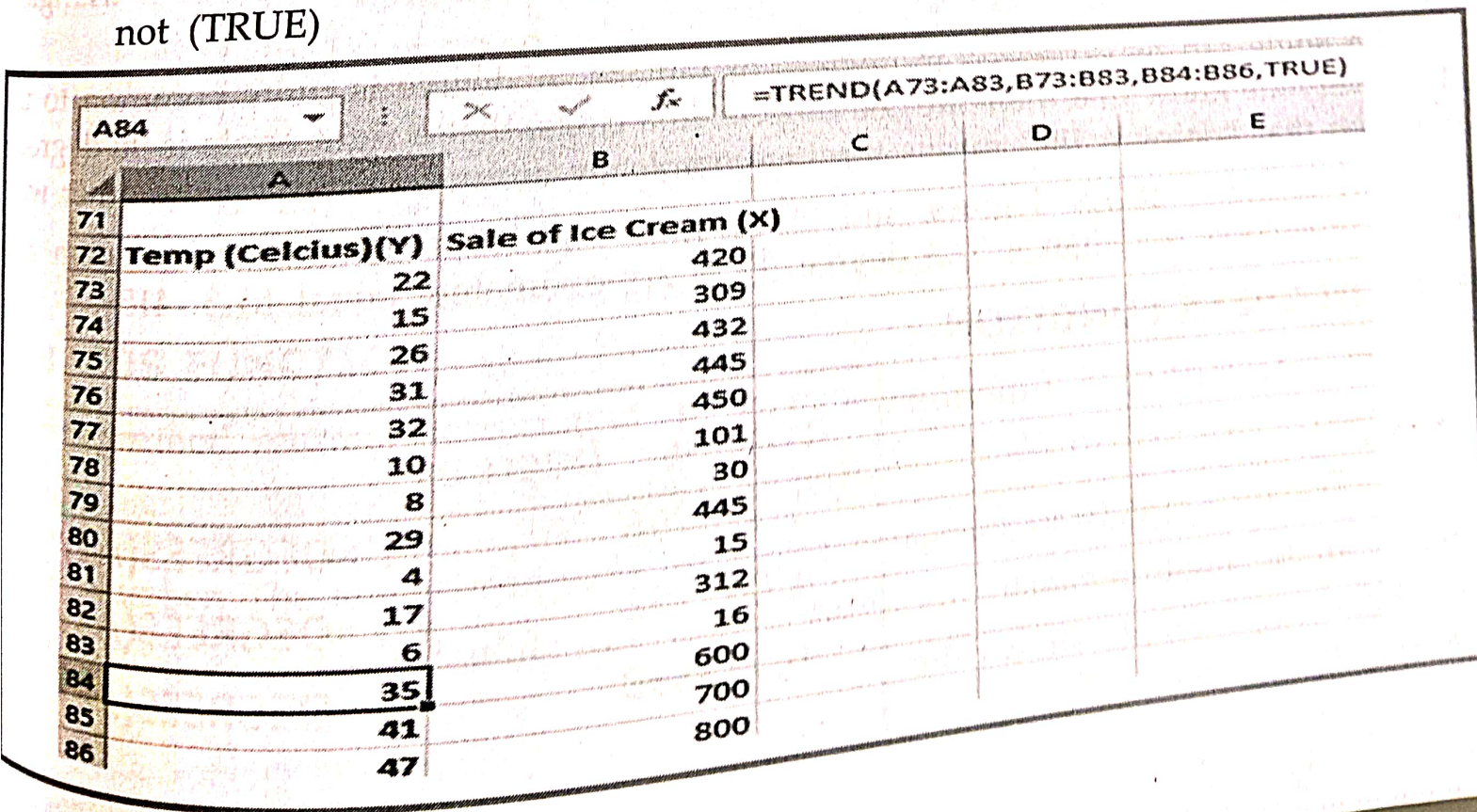
- **x** The data point of the value you want to predict.
- **known_y's** The known values for the known data points; the dependent variable.
- **known_x's** The known data points—typically increments of time (years, quarters, months); the independent variable.

In the following figure, it predicts the temperature as 35 if sale of the ice cream needed to be 600

Using the least squares method, the **TREND** function determines the values that plot a straight line based on a data set. In Figure 11.50, the trend values have been calculated. The known_y and known_x values are based on the formula y=m*x+b. Because the new y values (the values that the function returns) are along the linear trendline that best fits your data. TREND returns the y-values along a linear trendline that best fits the values in a data set.

TREND(known_y's, known_x's, new_x's, const)

- **known_y's** The known values for the known data points; the dependent variable.
- **known_x's** The known data points. If omitted assumed to be the array {1, 2, 3...} the same size as the known_y's.
- **new_x's** Are the new x-values for which you want TREND to return corresponding y-values. If omitted assumed to be the same as the known_x's.
- **const** TRUE or FALSE indicating whether the y-intercept (b) is = 0 (FALSE) or not (TRUE)



Because this function returns an array:

- The result range must be selected before you begin the function.
- The formula must be array entered by using Ctrl+Shift+Enter, instead of just Enter.

The FORECAST() and TREND() functions predict missing values based on existing data. These "missing" values may represent experimental data you didn't collect, or future sales numbers for dates that haven't yet occurred. Excel examines the data you have so far (the more the better), and uses a mathematical technique known as the method of least squares to generate a trend line. Excel then uses this trend line to generate the other values you requested. Both FORECAST() and TREND() work similarly and produce the same answers. However, FORECAST() calculates just a single value; TREND(), on the other hand, is an array function that gives you multiple values.

To predict a new value using the FORECAST() function, for instance, you supply the function with a new X value. It then uses the known relationship between the X and Y values to guess the corresponding new Y value.

You can also use the TREND() function to generate expected sales based on multiple inquiry numbers (as opposed to the single sales value that FORECAST() generates). Using the previous example, you could theoretically use TREND() to predict sales for the next six months.

8.5 TRIGONOMETRIC FUNCTIONS

Excel's trigonometric functions can be useful in a wide range of applications. From creating graphs for your math class to calculating complex financial models, Excel makes it easy to work with angles.

The sine, cosine, and tangent of an angle are likely the most used values in trigonometry calculations. They provide answers about the relationships of a triangle's angles to the sides of the triangle.

An angle can be expressed in degrees or radians. A degree is more common to us. Most everyone know that there are 360 degrees in a circle, that a right angle is 90 degrees, and even that "doing a 180" means turning completely around and going the other way.

One radian = 180/pi degrees, and one degree = pi/180 radians. For lowdown and quick conversion: 1 radian = 57.3 degree. Pi is approximately equal to 3.14159.

8.5.1 DEGREES FUNCTION

The DEGREES function converts radians into degree.

DEGREES(angle)

angle: angle in radian

Following example shows the working of function

Radians	Degrees	
1.570796	90	=DEGREES(A2)
1.047198	60	=DEGREES(A3)
0.785398	45	=DEGREES(A4)

8.5.2 RADIANS FUNCTION

The RADIANS function converts degree into radians.

RADIANS(angle)

	A	B	C
	Degree	Radians	
	90	1.570796	=RADIANS(A2)
	60	1.047198	=RADIANS(A3)
	45	0.785398	=RADIANS(A4)

- **angle:** angle in degree

8.5.3. SIN FUNCTION

This method returns the sine of the given angle

SIN(number)

- **Number:** The angle in radians for which you want the sine

Following example calculate the sine value correspondence to given radian in the previous section

	A	B	C
1	Radians	SIINE	
2	1.5708	1	=SIN(A2)
3	1.0472	0.866025628	=SIN(A3)
4	0.7854	0.707106666	=SIN(A4)

8.5.4. COS FUNCTION

This method returns the cosine of the given angle

COS(number)

- **Number:** The angle in radians for which you want the cosine

Following example calculate the cosine value correspondence to given radian values.

	A	B	C
1	Radians	COSIINE	
2	1.5708	3.26795E-07	=COS(A2)
3	1.0472	0.499999611	=COS(A3)
4	0.7854	0.707106897	=COS(A4)
5			

8.5.3. TAN FUNCTION

This method returns the tangent of the given angle

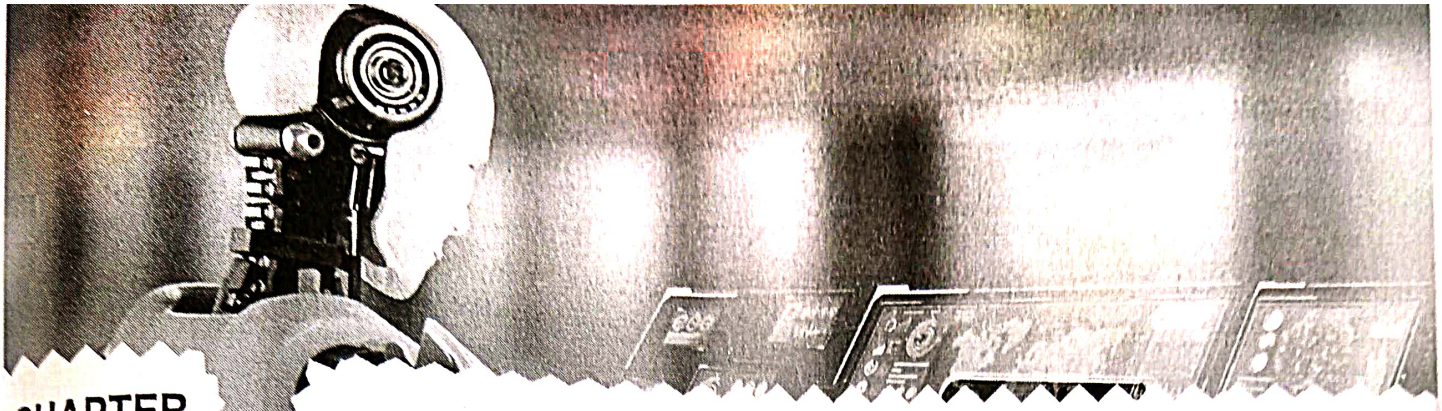
TAN(number)

- **Number:** The angle in radians for which you want the tangent

Following example calculate the tangent value correspondence to given radian valu

	A	B	C
1	Radians	TANGENT	
2	1.5708	3060023.307	=TAN(A2)
3	1.0472	1.732052603	=TAN(A3)
4	0.7854	0.999999673	=TAN(A4)

☆☆☆☆☆



CHAPTER

9 STRING, DATE & TIME AND LOGICAL EXCEL FUNCTION

CHAPTER AT A GLANCE

9.1 String/Text Functions	9.3 Logical Functions
9.2 Date and Time Functions	9.4 Database Functions

9.1 STRING/TEXT FUNCTIONS

In Excel, text is any collection of alphanumeric characters that isn't a numeric value, a date or time value, or a formula. Words, names, and labels are all obviously text values, but so are cell values preceded by an apostrophe (') or formatted as text. Text values are also called string.

Every character you can display on your screen has its own underlying numeric code. For example, the code for the uppercase letter A is 65, whereas the code for the ampersand (&) is 38. These codes apply not only to the alphanumeric characters accessible via your keyboard but also to extra characters. The collection of these characters is called the ANSI character set, and the numbers assigned to each character are called the ANSI codes

9.1.1. CHAR FUNCTION

Excel enables you to determine the character represented by an ANSI code using the CHAR()function:

CHAR(number)

- **number** The ANSI code, which must be a number between 1 and 255

For example, the following formula displays the copyright symbol (ANSI code 169):

CHAR(169)

9.1.2. CODE FUNCTION

The CODE() function is the opposite of CHAR(). That is, given a text character, CODE() returns its ANSI code value:

CODE(text)

- text A character or text string. Note that if you enter a multicharacter string, CODE() returns the ANSI code of the first character in the string.

For example, the following formulas both return 83, the ANSI code of the uppercase letter S:

CODE("S")

CODE("Space")

9.1.3. CONCATENATE FUNCTION AND OPERATOR

The &is an operator that you use to join text. Say that you have a worksheet with first name in Column A and last name in Column B, as shown in Figure. You need to put these names together in a single cell. If you use the formula =A2&B2 in Cell C2, Excel smashes the names together (for example, Gurdas Singh).

C2 | =A2&" "&B2

	A	B	C
1	First Name	Last Name	Name
2	Asha	Rani	Asha Rani
3	Gurdas	Singh	Gurdas Singh
4	Balkar	Singh	Balkar Singh
5	Devki	Dabby	Devki Dabby
6	Ramanpreet	Kaur	Ramanpreet Kaur

Instead, you have to join three elements. In between A2 and B2, you need to join a single space in double quotes. The formula to do this is =A2&" "&B2.

CONCATENATE (text1, text2, ...)

The CONCATENATE function joins several text strings into one text string. The arguments text1, text2, ... are 1 to 30 text items to be joined into a single text item. The text items can be text strings, numbers, or single-cell references.

The problem with this function is that it can only select single cell references. An attempt to use =CONCATENATE(A2:B2) returns a #VALUE! error. If you have to enter =CONCATENATE(A2, " ", B2), it is easier to use =A2&" "&B2. Further, the function can handle only 30 references.

C2 =CONCATENATE(A2," ",B2)

	A	B	C
1	First Name	Last Name	Name
2	Asha	Rani	Asha Rani
3	Gurdas	Singh	Gurdas Singh
4	Balkar	Singh	Balkar Singh
5	Devki	Dabby	Devki Dabby
6	Ramanpreet	Kaur	Ramanpreet Kaur

If you are joining cells with spaces in between, you will run out of terms after just 15 cells.

With the &, you can join more than 30 items.

9.1.4. UPPER, LOWER AND PROPER FUNCTION

The UPPER() function converts a specified string to all-uppercase letters:

UPPER(text)

- **text** The string you want to convert to uppercase

The LOWER() function converts a specified string to all-lowercase letters:

LOWER(text)

- **text** The string you want to convert to lowercase

The PROPER() function converts a specified string to proper case, which means the first letter of each word appears in uppercase and the rest of the letters appear in lowercase:

PROPER(text)

- **text** The string you want to convert to proper case

=UPPER(C2)

C	D	E	F
Name	UPPER	LOWER	PROPER
Asha rani	ASHA RANI	asha rani	Asha Rani
Gurdas singh	GURDAS SINGH	gurdas singh	Gurdas Singh
Balkar singh	BALKAR SINGH	balkar singh	Balkar Singh
Devki dabby	DEVKI DABBY	devki dabby	Devki Dabby
Ramanpreet kaur	RAMANPREET KAUR	ramanpreet	Ramanpreet Kaur

9.1.5. LEFT, RIGHT A MID FUNCTIONS

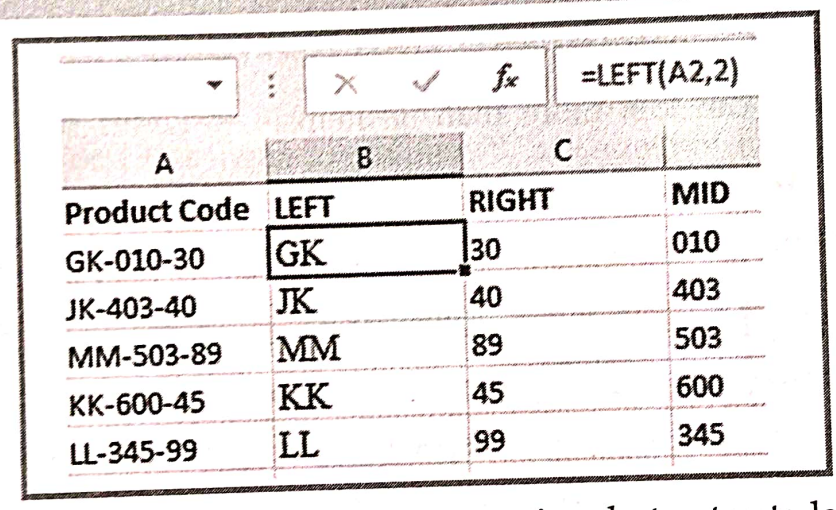
The LEFT(), MID(), and RIGHT() functions let you extract a portion of text from a larger text string. For example, the LEFT() function takes two arguments: the text you want to examine, and the number of characters that Excel should extract, starting from the string's left side:

LEFT(text, num_characters)

- **text** The string from which you want to extract the substring
- **num_chars** The number of characters you want to extract from the left

To take the first two letters from the text in cell A2, you'd use the formula:

LEFT(A2, 2)



	A	B	C	
	Product Code	LEFT	RIGHT	MID
	GK-010-30	GK	30	010
	JK-403-40	JK	40	403
	MM-503-89	MM	89	503
	KK-600-45	KK	45	600
	LL-345-99	LL	99	345

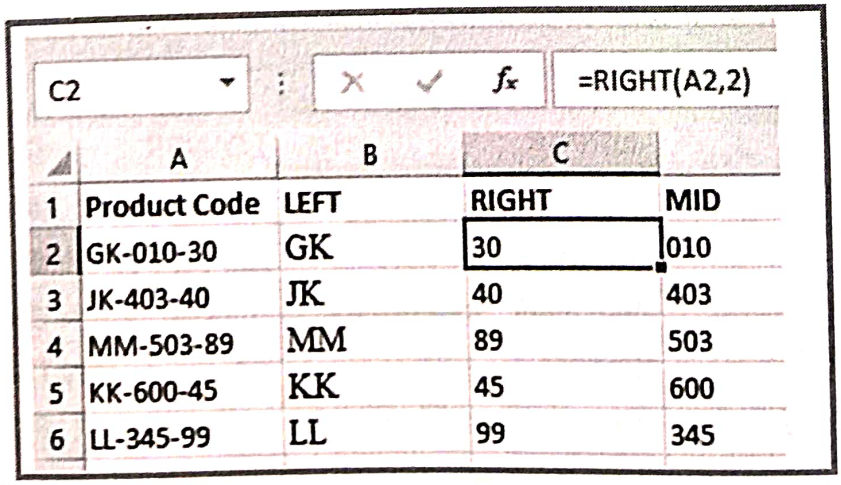
The RIGHT() function performs the same operation, but extracts letters starting from the right side of the string.

RIGHT(text [, num_chars])

- **text** The string from which you want to extract the substring
- **num_chars** The number of characters you want to extract from the left

To take the last two letters from the text in cell A2, you'd use the formula:

=RIGHT(A2, 2)

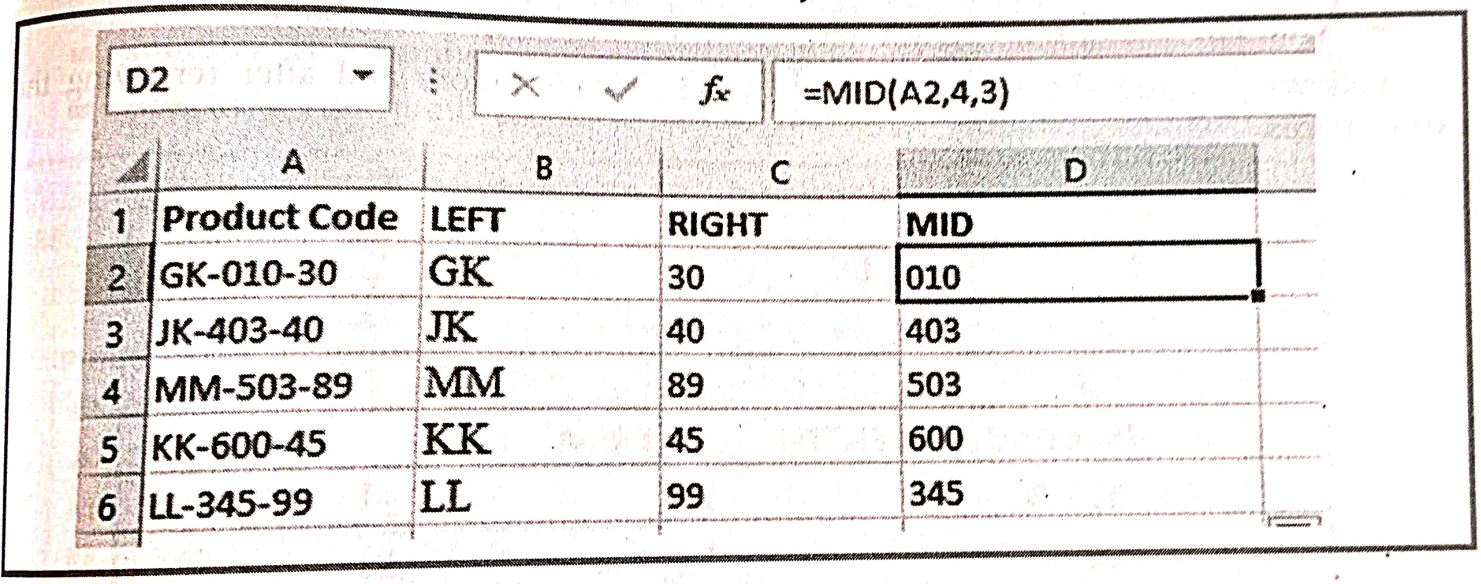


	A	B	C	
1	Product Code	LEFT	RIGHT	MID
2	GK-010-30	GK	30	010
3	JK-403-40	JK	40	403
4	MM-503-89	MM	89	503
5	KK-600-45	KK	45	600
6	LL-345-99	LL	99	345

The MID() function is more powerful than the LEFT() and RIGHT() functions, as it has the ability to extract a consecutive series of characters from anywhere inside a string. When using the MID() function, you need to supply three arguments: the text you're evaluating, the starting position of the extraction, and the number of characters you want to retrieve. Excel numbers each letter and space in a string, starting with 1 for the first character, 2 for the second character, and so on. That means that if you specify a starting position of 3 and a length of 2, Excel extracts the third and fourth characters from a string. The basic formula looks like this:

MID(text, start_num, num_chars)

- **text** The string from which you want to extract the substring
- **start_num** The character position at which you want to start extracting the substring
- **num_chars** The number of characters you want to extract

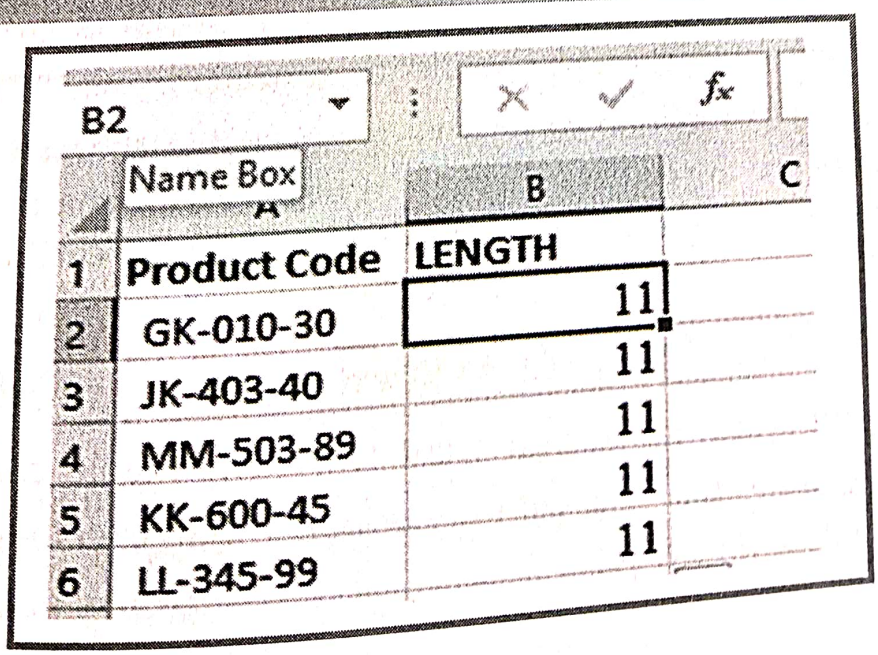


	A	B	C	D
1	Product Code	LEFT	RIGHT	MID
2	GK-010-30	GK	30	010
3	JK-403-40	JK	40	403
4	MM-503-89	MM	89	503
5	KK-600-45	KK	45	600
6	LL-345-99	LL	99	345

9.1.6. LENGTH FUNCTION

LEN()—short for LENgth—counts the number of characters in a string of text.

LEN(text)



	A	B	C
1	Product Code	LENGTH	
2	GK-010-30	11	
3	JK-403-40	11	
4	MM-503-89	11	
5	KK-600-45	11	
6	LL-345-99	11	

- **text** The string from which you want to count the number of characters

For example, the result of the following formula is 5. LEN("Hello")

9.1.7. TRIM AND CLEAN FUNCTIONS

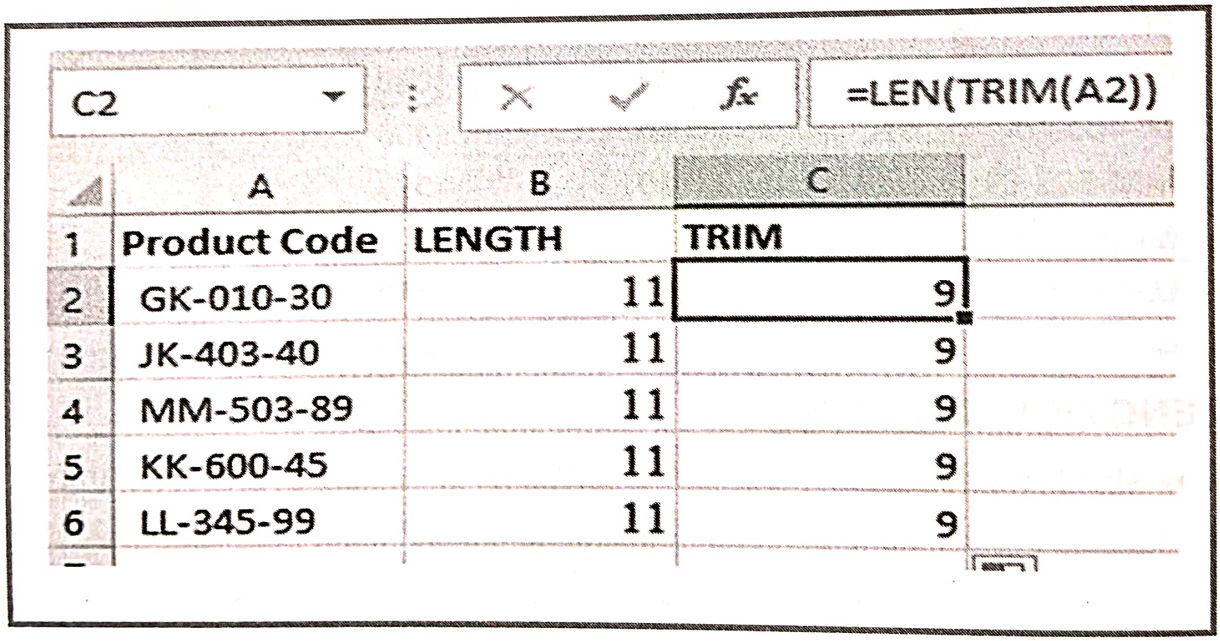
The TRIM function removes all spaces from text except for single spaces between words.

You use TRIM on text that you have received from another application that may have irregular spacing. Here, irregular means all spaces before and after the string, as well as two or more consecutive spaces within the string. In the latter case, TRIM() removes all but one of the consecutive spaces. The argument *text* is the text from which you want spaces removed

TRIM(text)

- **text** The string from which you want the excess spaces removed

Following figure shows the length of product code before and after removing the extra spaces.



	A	B	C
1	Product Code	LENGTH	TRIM
2	GK-010-30	11	9
3	JK-403-40	11	9
4	MM-503-89	11	9
5	KK-600-45	11	9
6	LL-345-99	11	9

You use the CLEAN() function to remove nonprintable characters from a string. the nonprintable characters are the codes 1 through 31 of the ANSI character set. The CLEAN() function is most often used to remove line feeds (ANSI 10) or carriage returns (ANSI 13) from multiline data.

CLEAN(text)

- **text** The string from which you want the nonprintable characters removed

Following figures show the data with nonprintable characters and then clean function remove the nonprintable characters.

IMPORTED VALUE		IMPORTED VALUE	CLEANED VALUE
254686988	=CLEAN(C9)	254686988	254686986
254688 87		254688 87	254686987
254686988		254686988	254686988
2546 8989		2546 8989	254686989
2546 6990		2546 6990	254686990

9.1.8. DOLLAR AND FIXED FUNCTION

The DOLLAR() function converts a numeric value into a string value that uses the Currency format:

DOLLAR(number [, decimals])

- **number** The number you want to convert
- **decimals** The number of decimals to display

B8 ="The total sale for this quarter is " & DOLLAR(B7, 0)

	A	B	C	D	E
1	Product Code	Amouont			
2	GK-010-30	150			
3	JK-403-40	200			
4	MM-503-89	450			
5	KK-600-45	600			
6	LL-345-99	789			
7		2189			
8		The total sale for this quarter is $2,189			

For some kinds of numbers, you can control the number of decimals and whether commas are inserted as the thousands separator by using the FIXED() function:

FIXED(number [, decimals] [, no_commas])

- **number** The number you want to convert to a string.
- **decimals** The number of decimals to display.
- **no_commas** A logical value that determines whether commas are inserted into the string. Use TRUE to suppress commas; use FALSE to include commas.

For example, the following formula uses the SUM() function to take a sum of a range and applies the FIXED() function to the result so that it's displayed as a string with commas and no decimal places:

B8 ⋮ ✕ ✓ f_x ="Total sale: " & FIXED(SUM(B2:B6), 0, FALSE) & " of parts."

	A	B	C	D	E	F
1	Product Code	Amouont				
2	GK-010-30	150				
3	JK-403-40	200				
4	MM-503-89	450				
5	KK-600-45	600				
6	LL-345-99	789				
7		2189				
8		Total sale: 2,189 of parts.				

9.1.9. REPT FUNCTION

The REPT function repeats text a given number of times. You use REPT to fill a cell with a number of instances of a text string. This function takes the following arguments:

REPT(text, number_times)

- **text** This is the text you want to repeat.
- **number_times** This is a positive number that specifies the number of times to repeat text. If number_times is 0, REPT returns "" (that is, empty text). If number_times is not an integer, it is truncated. The result of the REPT function cannot be longer than 32, 767 characters.

B2 ⋮ ✕ ✓ f_x =A2 & REPT(".",15)

	A	B	C
1	Chapters	With Rept	
2	Chapter 1	Chapter 1.....	
3	Chapter 2	Chapter 2.....	
4	Chapter 3	Chapter 3.....	
5	Chapter 4	Chapter 4.....	
6	Chapter 5	Chapter 5.....	
7	Chapter 6	Chapter 6.....	

9.1.10. EXACT FUNCTION

The EXACT function compares two text strings and returns TRUE if they are exactly the same and FALSE otherwise. EXACT is case-sensitive but ignores formatting differences. You use EXACT to test text being entered into a document. This function takes the following arguments:

EXACT(text1, text2)

- **text1** This is the first text string.
- **text2** This is the second text string

C2 =EXACT(A2,B2)

	A	B	C	D
1	Text1	Text2	Reuslt	
2	Office	Office	TRUE	
3	OFFICE	OffiCe	FALSE	
4	office	office	TRUE	
5	Office	officE	FALSE	

9.1.11. FIND AND SEARCH FUNCTION

FIND finds one text string (find_text) within another text string (within_text) and returns the number of the starting position of find_text from the first character of within_text. You can also use SEARCH to find one text string within another, but unlike SEARCH, FIND is case-sensitive and doesn't allow wildcard characters. The FIND function takes the following arguments:

FIND(find_text, within_text, start_num)

- **find_text** This is the text you want to find. If find_text is "" (that is, empty text), FIND matches the first character in the search string (that is, the character numbered start_num or 1). find_text cannot contain any wildcard characters.
- **within_text** This is the text that contains the text you want to find.
- **start_num This** specifies the character at which to start the search. The first character in within_text is character number 1. If you omit start_num, it is assumed to be 1.

If find_text does not appear in within_text, FIND return a #VALUE! error. If start_num is not greater than zero, FIND returns a #VALUE! error. If start_num is greater than the length of within_text, FIND return a #VALUE! error.

C2 =FIND(B2,A2,1)

	A	B	C
1	Sourec Text	Search Text	Position
2	Microsoft	soft	6
3	Microsoft Windows	Windows	11
4	Linux windows	Windows	#VALUE!
5	Mac OS	ac	2

In the following figure find function start from position 1 and find the text but in row 4 has 'windows' text in different cases so it displays the #value!.

SEARCH returns the number of the character at which a specific character or text string is first found, beginning with start_num. You use SEARCH to determine the location of a character or text string within another text string so that you can use the MID or REPLACE functions to change the text. The SEARCH function takes the following arguments:

SEARCH(find_text, within_text, start_num)

- **find_text** This is the text you want to find. You can use the wildcard characters question mark (?) and asterisk (*) in find_text. A question mark matches any single character; an asterisk matches any sequence of characters. If you want to find an actual question mark or asterisk, you type a tilde (~) before the character. If find_text is not found, a #VALUE! error is returned.
- **within_text**—This is the text in which you want to search for find_text.
- **start_num**—This is the character number in within_text at which you want to start searching. If start_num is omitted, it is assumed to be 1. If start_num is not greater than zero or is greater than the length of within_text, a #VALUE! error is returned.

But when search function was used then it displays all the positions.

C4 =SEARCH(B4,A4,1)

	A	B	C
1	Sourec Text	Search Text	Position
2	Microsoft	soft	6
3	Microsoft Windows	Windows	11
4	Linux windows	Windows	7
5	Mac OS	ac	2

9.1.12. SUBSTITUTE AND REPLACE FUNCTIONS

The SUBSTITUTE function substitutes new_text for old_text in a text string. You use SUBSTITUTE when you want to replace specific text in a text string; you use REPLACE when you want to replace any text that occurs in a specific location in a text string. The SUBSTITUTE function takes the following arguments:

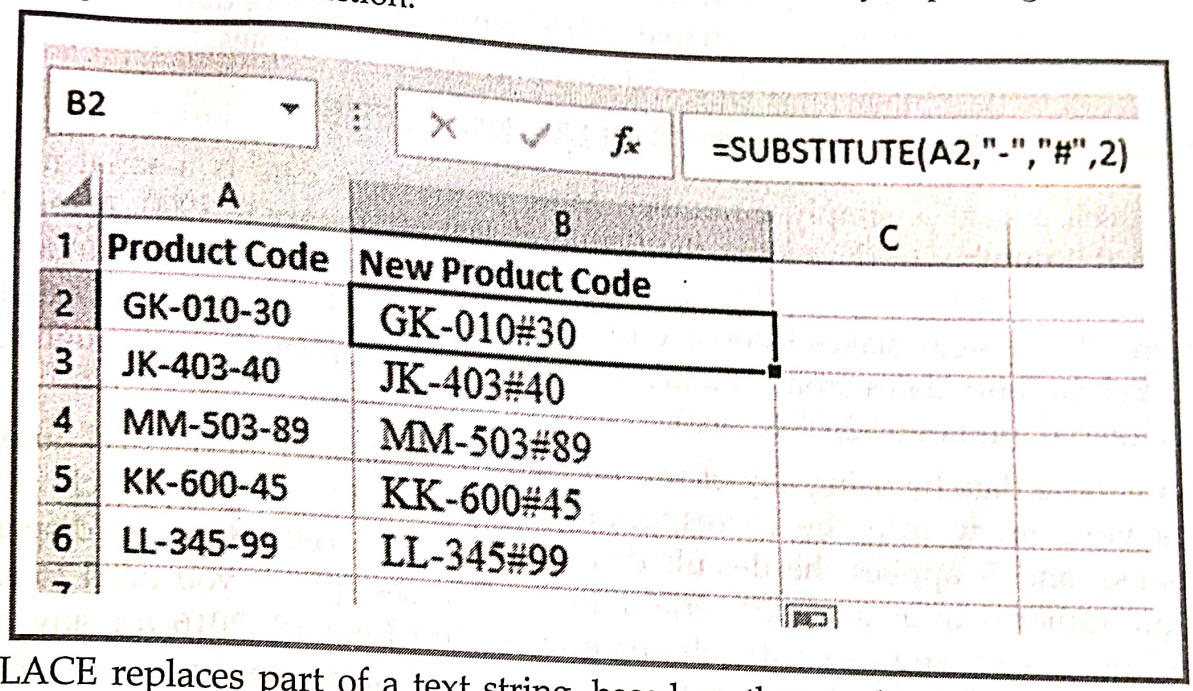
SUBSTITUTE(text, old_text, new_text, instance_num)

- **text** This is the text or the reference to a cell that contains text for which you want to substitute characters.
- **old_text** This is the text you want to replace.

- **new_text** This is the text you want to replace old_text with.
- **instance_num**—This specifies which occurrence of old_text you want to replace with new_text. If you specify instance_num, only that instance of old_text is replaced. Otherwise, every occurrence of old_text in text is changed to new_text.

For example, =SUBSTITUTE("Sales Data", "Sales", "Cost") would generate "Cost Data".

In the following figure new product code is generated by replacing the second '-' with '#' using substitute function.



	A	B	C
1	Product Code	New Product Code	
2	GK-010-30	GK-010#30	
3	JK-403-40	JK-403#40	
4	MM-503-89	MM-503#89	
5	KK-600-45	KK-600#45	
6	LL-345-99	LL-345#99	

REPLACE replaces part of a text string, based on the number of characters specified, with a different text string. This function takes the following arguments:

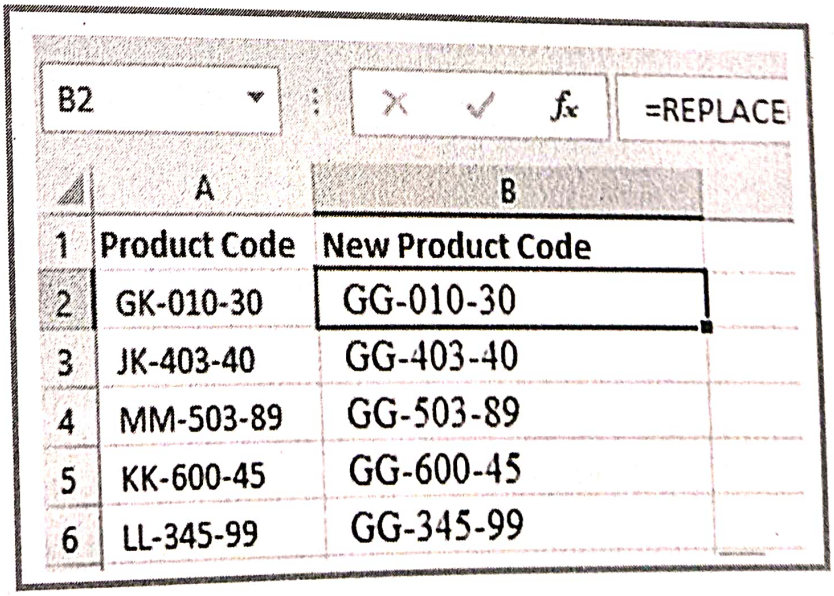
REPLACE(old_text, start_num, num_chars, new_text)

- **old_text This** is text in which you want to replace some characters.
- **start_num This** is the position of the character in old_text that you want to replace with new_text.
- **num_chars This** is the number of characters in old_text that you want REPLACE to replace with new_text.

new_text This is the text that will replace characters in old_text.

In the following figure new product code is generated by prefixing 'GG' and replacing the first two character of product code by entering the starting position as 3.

In order to successfully use REPLACE, you have to use functions



	A	B
1	Product Code	New Product Code
2	GK-010-30	GG-010-30
3	JK-403-40	GG-403-40
4	MM-503-89	GG-503-89
5	KK-600-45	GG-600-45
6	LL-345-99	GG-345-99

to determine the location and number of characters to replace. In most circumstances, SUBSTITUTE is easier to use.

9.2 DATE AND TIME FUNCTIONS

Excel's date and time functions enable you to convert dates and times to serial numbers and perform mathematical operations on those numbers. This capability is useful for such things as accounts receivable aging, project scheduling, time-management applications, and much more. This chapter introduces you to Excel's date and time functions and puts them through their paces with many practical examples.

9.2.1 UNDERSTANDING DATE SERIAL NUMBERS

To Excel, a date is simply a number. More precisely, a date is a serial number that represents the number of days since the fictitious date of January 0, 1900. A serial number of 1 corresponds to January 1, 1900; a serial number of 2 corresponds to January 2, 1900; and so on. This system makes t possible to create formulas that perform calculations with dates. Excel support dates from January 1, 1900 (serial number = 1), through December 31, 9999 (serial number =2, 958, 465).

You enter a date by using any of several recognized date formats. Excel automatically converts your entry into the corresponding date serial number (which it uses for calculations), and it applies the default date format to the cell so that it displays as an actual date rather than as a cryptic serial number. For example, if you need to enter June 18, 2016, into a cell, you can enter the date by typing June 18, 2016 (or any of several different date formats). Excel interprets your entry and stores the value 42539, the date serial number for that date. It also applies the default date format, so the cell contents may not appear exactly as you typed them.

When you activate a cell that contains a date, the Formula bar shows the cell contents formatted by using the default date format — which corresponds to your system's short date format. The Formula bar doesn't display the date's serial number. If you need to find out the serial number for a particular date, format the cell with a nondate number format.

Serial Number	Date
366	December 31, 1900
16229	June 6, 1944
43830	December 31, 2019

9.2.2 UNDERSTANDING TIME

Excel also stores times of day as numbers behind the scenes. In fact, Excel internally stores every time value as a fractional number from 0 to 1. The number 0 represents 12:00 AM (the start of the day) and 0.99999 represents 11:59:59 PM (the end of the day). Because Excel stores times as a single number, it's easy to subtract one time value from another. However, time values can have varying degrees of precision. If your time's accurate down to the millisecond, it includes more decimal places (up to eight). For example, Excel stores 10:30 AM as 0.4375, whereas it stores 10:30:32.34 as 0.437874306.

Serial Number	Time
0.25	6:00:00 a.m.
0.375	9:00:00 a.m.
0.70833	5:00:00 p.m.
.99999	11:59:59 p.m.

You can combine the two types of serial numbers. For example, 43830.5 represents noon on December 31, 2019. The advantage of using serial numbers in this way is that it makes Calculations involving dates and times very easy. A date or time is really just a number, so any mathematical operation you can perform on a number can also be performed on a date. This is invaluable for worksheets that track delivery times, monitor accounts receivable or accounts payable aging, calculate invoice discount dates, and so on.

9.2.3. NOW AND TODAY FUNCTION

Excel offers two functions for calculating the current date: NOW and TODAY.

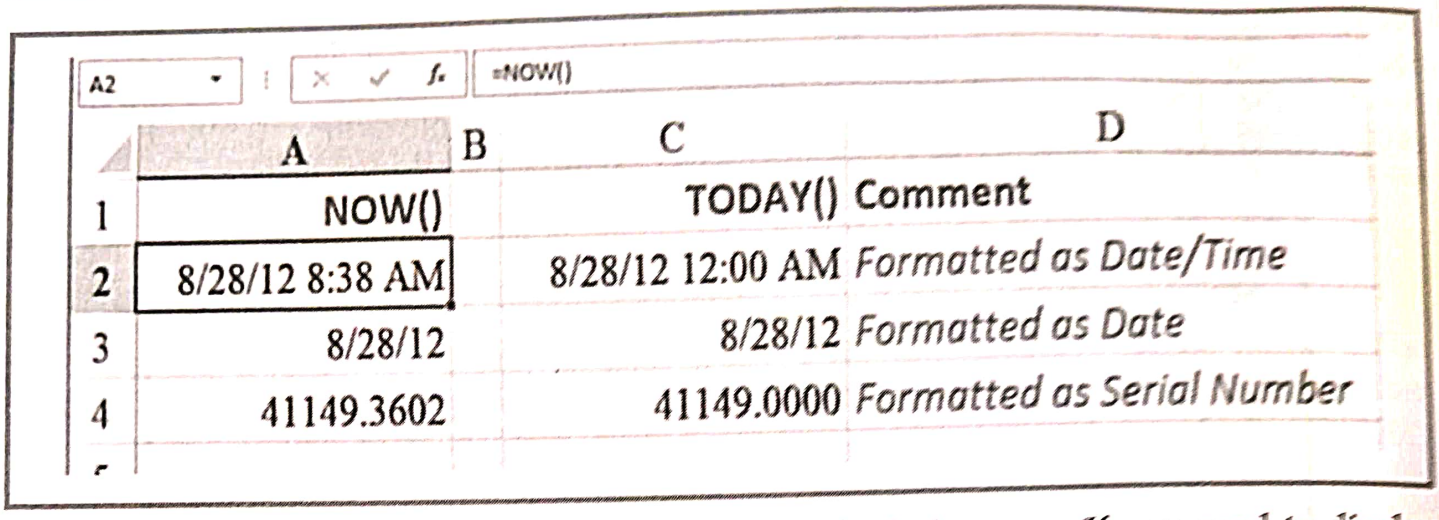
These functions are excellent for figuring out the number of days until a deadline or how late an open receivable might be.

NOW()

TODAY()

NOW returns the serial number of the current date and time. TODAY returns the serial number of the current date. The TODAY function returns today's date, without any time attached. The NOW function returns the current date and time.

Both of these functions can be made to display the current date, but there is an important distinction when you are performing calculations with the function. In the following figure column A contains NOW functions, and column C contains TODAY functions. Row 2 is formatted as a date and time. Row 3 is formatted as a date.



	A	B	C	D
1	NOW()		TODAY()	Comment
2	8/28/12 8:38 AM		8/28/12 12:00 AM	Formatted as Date/Time
3	8/28/12		8/28/12	Formatted as Date
4	41149.3602		41149.0000	Formatted as Serial Number

Row 4 is formatted as numeric. Cell A3 and C3 look the same. If you need to display the date without using it in a calculation, then NOW or TODAY work fine.

9.2.4. DATE FUNCTION

The DATE function returns the serial number that represents a particular date. The DATE function is amazing because it enables Excel to deal perfectly. For example, the following expression returns January 1, 2020:

```
DATE (2019, 12, 32)
```

Here, DATE() adds the extra day (there are 31 days in December) to return the date of the next day. Similarly, the following expression returns January 25, 2020:

DATE(2019, 13, 25) with invalid dates.

You can do that by using Excel's DATE() function:

```
DATE(year, month, day)
```

- **year** This argument can be one to four digits. If *year* is between 0 and 1899 (inclusive), Excel adds that value to 1900 to calculate the year. For example, =DATE(100, 1, 2) returns January 2, 2000 (1900+100). If year is between 1900 and 9999 (inclusive), Excel uses that value as the year. For example, =DATE(2000, 1, 2) returns January 2, 2000. If *year* is less than 0 or is 10000 or greater, Excel returns a #NUM! error.
- **month** This is a number representing the month of the year. If month is greater than 12, month adds that number of months to the first month in the year specified. For example, =DATE(1998, 14, 2) returns the serial number representing February 2, 1999.
- **day** This is a number representing the day of the month. If *day* is greater than the number of days in the month specified, it adds that number of days to the

first day in the month. For example, =DATE (1998, 1, 35) returns the serial number representing February 4, 1998. In a trivial example, =DATE(2011, 3, 5) returns March 5, 2011.

Following figure shows various results of the DATE function:

D2 : =DATE(A2,B2,C2)

	A	B	C	D
1	Year	Month	Day	DATE
2	2015	16	45	5/15/2016
3	2016	1	60	2/29/2016
4	2015	1	60	3/1/2015
5	2015	3	5	3/5/2015

9.2.5. DATEVALUE FUNCTION

If you have a date value in string form, you can convert it to a date serial number by using the DATEVALUE() function:

DATEVALUE(date_text)

- **date_text** The string containing the date

For example, the following expression returns the date serial number for the string August 23, 2019:

DATEVALUE("August 23, 2019")

B2 : =DATEVALUE("August 23, 2019")

	A	B	C	D	E	F	G
1							
2		43700					

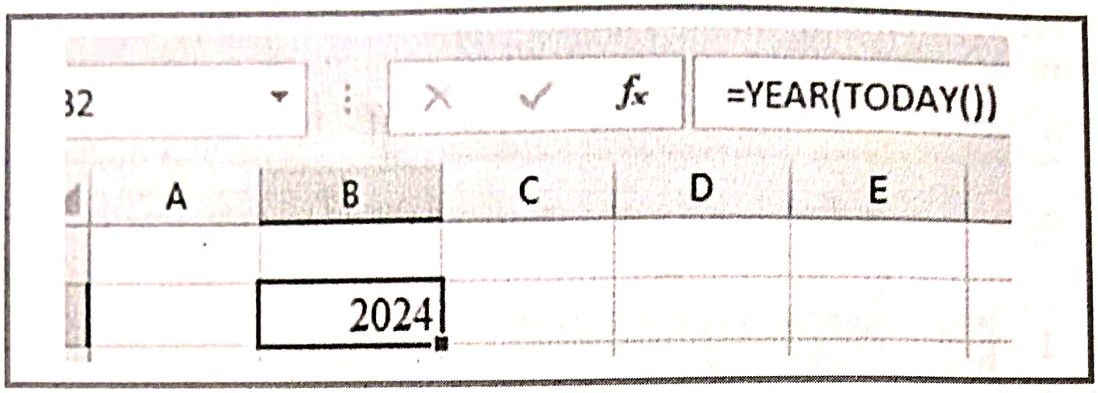
9.2.6. YEAR FUNCTION

The YEAR() function returns a four-digit number that corresponds to the year component of a specified date:

YEAR(date)

- **date** The date (or a string representation of the date) you want to work with

For example, if today is July 23, 2024, the following expression will return 2024:

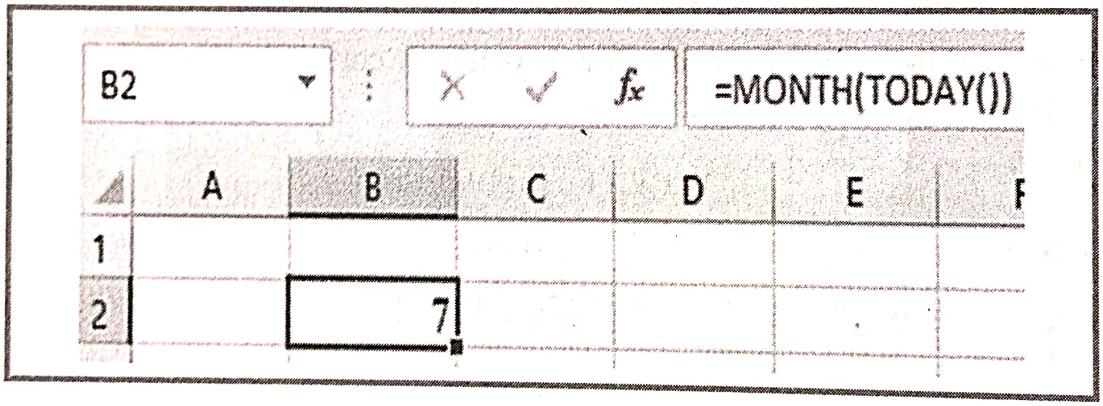


9.2.7. MONTH FUNCTION

The MONTH() function returns a number between 1 and 12 that corresponds to the month component of a specified date:

MONTH(date)

- **date** The date (or a string representation of the date) you want to work with.



9.2.8. DAY FUNCTION

The DAY() function returns a number between 1 and 31 that corresponds to the day component of a specified date:

DAY(date)

- **date** The date (or a string representation of the date) you want to work with

9.2.9. WEEKDAY FUNCTION

The WEEKDAY() function returns a number that corresponds to the day of the week upon which a specified date falls:

WEEKDAY(date [, return_type])

- **date** The date (or a string representation of the date) you want to work with.
- **return_type** An integer that determines how the value returned by WEEKDAY()corresponds to the days of the week:

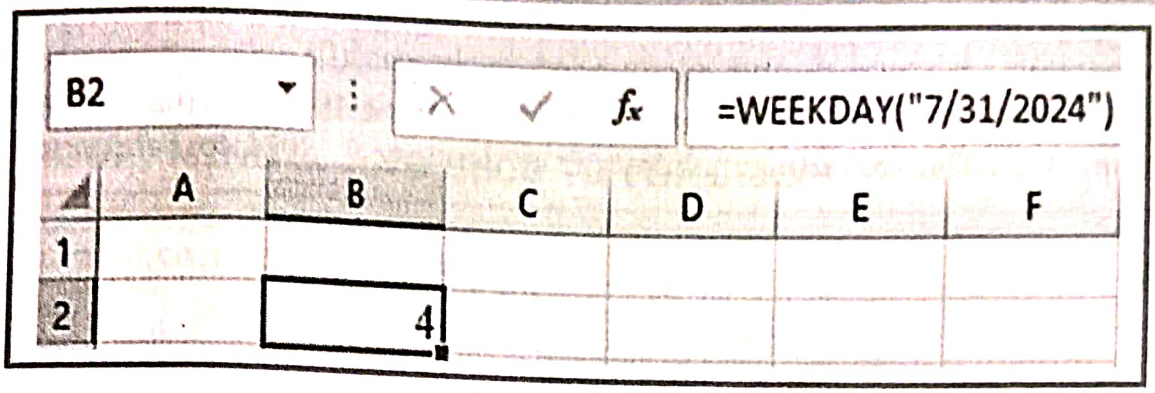
  1: The return values are 1 (Sunday) through 7 (Saturday); this is the default.

2: The return values are 1 (Monday) through 7 (Sunday).

3: The return values are 0 (Monday) through 6 (Sunday).

For example, the following expression returns 4 because July 31, 2024, is a Wednesday:

```
WEEKDAY("7/31/2024")
```



9.2.10. WEEKNUM FUNCTION

The WEEKNUM() function returns a number that corresponds to where the week that includes a specified date falls numerically during the year:

```
WEEKNUM(date [, return_type])
```

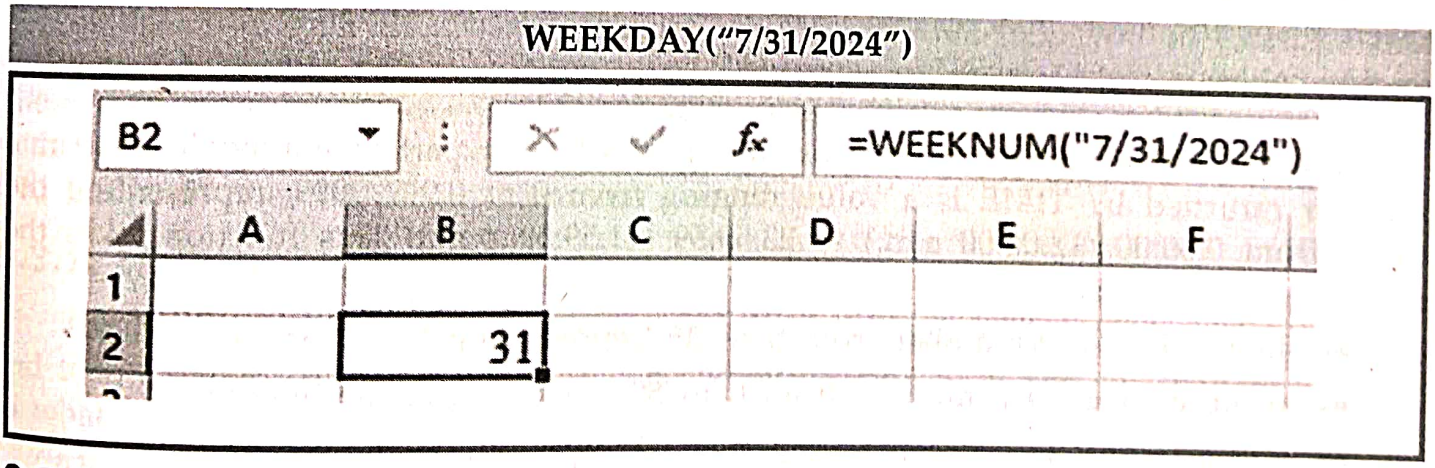
- **date** The date (or a string representation of the date) you want to work with.
- **return_type** An integer that determines how WEEKNUM() interprets the start of the week:

  1: The week begins on Sunday; this is the default.

  2: The week begins on Monday.

For example, the following expression returns 31 because July 31, 2024, is a Wednesday:

```
WEEKDAY("7/31/2024")
```



9.2.11. WORKDAYS FUNCTION

The WORKDAY() function, returns a date that is a specified number of working days from some starting date:Click here to view code image

```
WORKDAY(start_date, days [, holidays])
```

- **start_date** The original date (or a string representation of the date).
- **days** The number of workdays before or after start_date. Use a positive number to return a later date; use a negative number to return an earlier date. Non-integer values are truncated (that is, the decimal part is ignored).
- **holidays** A list of dates to exclude from the calculation. This can be a range of dates or an array constant—that is, a series of date serial numbers or date strings, separated by commas and surrounded by braces ({}).

Here's another expression that returns the date April 19, 2024, that is 65 workdays from January 17, 2024, excluding holidays from January 1, 2024 to February 16, 2024..

B4 =WORKDAY(B1,B2,E2:E11)

	A	B	C	D	E
1	Start Date	1/17/2024			Holidays
2	Work Day:	65			1/1/2024
3	End Date	Tuesday, July 21,2024			1/19/2024
4		4/19/2024			2/16/2024
5					5/25/2024
6					7/3/2024
7					9/7/2024
8					10/12/2024
9					11/11/2024
10					11/26/2024
11					12/25/2024

9.2.12. TIME FUNCTION

The TIME function is similar to the DATE function. It calculates a time serial number given a specific hour, minute, and second.

TIME(hour, minute, second)

The TIME function returns the decimal number for a particular time. The decimal number returned by TIME is a value ranging from 0 to 0.99999999, representing the times from 0:00:00 (12:00:00 a.m.) to 23:59:59 (11:59:59 p.m.). This function takes the following arguments:

- **hour**—This is a number from 0 to 23, representing the hour.
- **minute**—This is a number from 0 to 59, representing the minute.
- **second**—This is a number from 0 to 59, representing the second

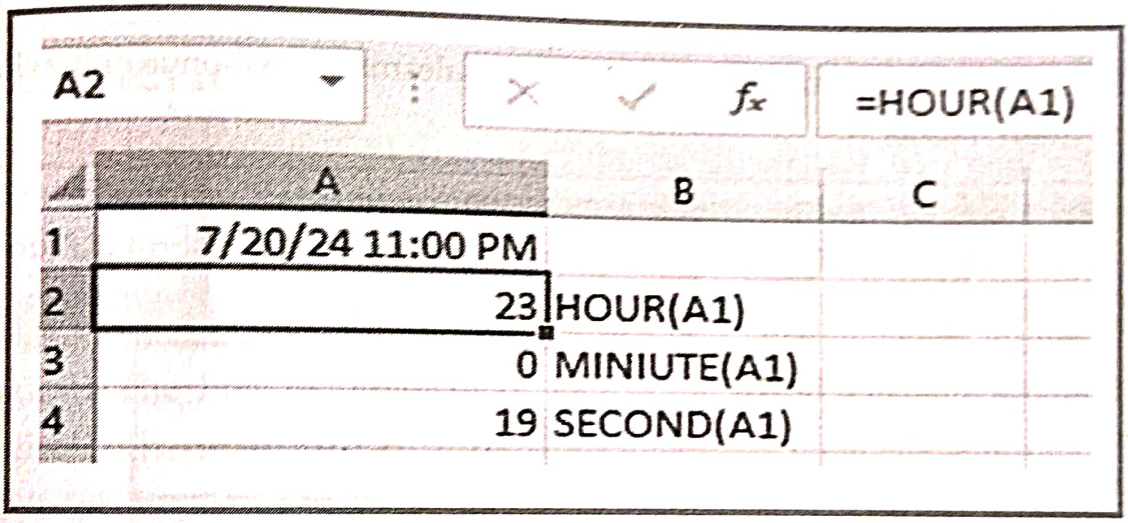
As with the DATE function, Excel can handle situations in which the minute or second argument calculates to more than 60. For example,

=TIME(12, 72, 120) evaluates to 1:14 PM.

	Hour	Minute	Second	TIME
8				
9	1	12	23	1:12:23 AM
10	13	12	23	1:12:23 PM
11	12	72	23	1:12:23 PM
12	37	12	23	1:12:23 PM

9.2.13. HOUR, MINUTE AND SECOND FUNCTION

- HOUR(*date*) returns the hour, from 0 to 23.
- MINUTE(*date*) returns the minute, from 0 to 59.
- SECOND(*date*) returns the second, from 0 to 59.



9.3 LOGICAL FUNCTIONS

Logical functions are decision-making instruments that help you to examine the contents of a cell, perform a calculation, and then compare the result to a specified value. The logical function may also be used to determine whether a condition is true or false, with the conclusions being used to display data and perform other calculations. IF, NOT, OR, IFS, and other logical functions are examples.

9.3.1. IF FUNCTION

The IF function is a check function that takes one value if the statement is Fulfilled and then another value if it is FALSE. The whole feature helps you to compare a value to what you predict in a rational manner.

The following arguments are used by the IF function.

IF (Logical_text, [Value_if_true], [Value_if_false])

- **Logical_text** That is the value or logical expression which are to be tested and evaluated as either TRUE or FALSE

- **Value_if_true** This is the value that will be returned if the logical test evaluates to TRUE
- **Value_if_false** This is the value that will be returned if the logical test evaluates to FALSE

The preceding logical operators can be used with this function:

Equal to (=)

Greater than (>)

Greater than or equal to (≥)

Less than (<)

Less than or equal to (£)

Not equalt (≠)

In the following figure Customer sales data. A salesman gets bonus only when monthly sale is more than 74000.

C3 =IF(B3 > 74000, "Got Bonus", "No Bonus")

	A	B	C	D	E
1	**Customers Sales Data**				
2	Sales Man	Sales	AND		
3	Raman	50000	No Bonus		
4	Amrit	65000	No Bonus		
5	Pawan	75000	Got Bonus		
6	Guri	85000	Got Bonus		
7	Jasmine	95000	Got Bonus		
8	Harman	120000	Got Bonus		

9.3.2. AND FUNCTION

The AND function is used to decide whether the conditions specified in a data set are TRUE, and it returns a FALSE value if all of the conditions are not satisfied; for example, B1 is greater than 50 and less than 100. The AND function is called with the following arguments.

AND (logical1, [logical2},)

- **Logical1** The first requirement or logical value to be evaluated is
- **Logical2** The second state or logical value to be evaluated

Following figure shows the working of AND function

C3 =AND(B3 > 48000, B3 < 89000)

	A	B	C
1	Customers Sales Data		
2	Sales Man	Sales	AND
3	Raman	50000	TRUE
4	Amrit	65000	TRUE
5	Pawan	75000	TRUE
6	Guri	85000	TRUE
7	Jasmine	95000	FALSE
8	Harman	120000	FALSE

9.3.3. OR FUNCTION

The OR function returns TRUE if all of the conditions are true and FALSE if any of the conditions are false. If one of the conditions is false, unlike the AND function, it'll be returned as FALSE. The arguments to the OR function are as follows:

OR (logical1, [logical2] ...)

- **logical1** The first requirement or logical value to determine
- **logical2** The second prerequisite or logical validity to determine.

In the following figure we use the OR instead of AND using previous data.

C4 =OR(B4 > 48000, B4 < 89000)

	A	B	C
1	Customers Sales Data		
2	Sales Man	Sales	AND
3	Raman	50000	TRUE
4	Amrit	65000	TRUE
5	Pawan	75000	TRUE
6	Guri	85000	TRUE
7	Jasmine	95000	TRUE
8	Harman	120000	TRUE

9.3.4. XOR FUNCTION

The XOR function checks whether an odd number of the arguments are TRUE.

XOR(logical1, logical2, ...)

The arguments logical1, logical2, ... are one to 254 conditions you want to test that can either be TRUE or FALSE. They can be logical values, arrays, or references.

9.3.5. NOT FUNCTION

In the language of Boolean logic, there are typically NAND, NOR, and XOR functions, which stand for Not And, Not Or, and Exclusive Or. To simplify matters, Excel offers the NOT function.

NOT(logical)

Quite simply, NOT reverses a logical value. TRUE becomes FALSE, and FALSE becomes TRUE when processed through a NOT function.

For example, suppose you need to find all flights landing outside Gujrat and Delhi. You can build a massive OR statement to find every airport code in the India. Alternatively, you can build an OR function to find Gujrat and Delhi and then use a NOT function to reverse the result:

=NOT(OR(A2="Gujrat", A2="Delhi")).

9.4 DATABASE FUNCTIONS

Excel includes functions exclusively for use with long tables. These are the database functions, a set of 11 functions that let you analyze groups of data. The database functions are similar to basic statistical functions like SUM(), AVERAGE(), and COUNT(). In fact, the database functions have the exact same names, but with an initial letter D—so you find a DSUM(), DAVERAGE(), DCOUNT(), and so on. The database functions differ from their non-database counterparts in that they can selectively filter out rows. In other words, when you use DSUM(), you can specify a set of criteria that a record must match in order to be included in the sum.

The database functions, regardless of the difference in and the computations that they perform, all take the same three arguments as illustrated by the DAVERAGE function:

DAVERAGE(database, field, criteria)

The arguments for the database functions require the following information:

- **database** is the argument that specifies the range containing the list and it must include the row of field names in the top row.
- **field** is the argument that specifies the field whose values are to be calculated by the database function (averaged in the case of the DAVERAGE function). You can specify this argument by enclosing the name of field in double quotes (as in "Salary" or "Date Hired"), or you can do this by entering the number of the column in the data list (counting from left to right with the first field counted as 1). Note that in specifying the field argument, you must refer to a column in the data list that contains numeric or date data for all the database functions with the

exception of DGET. All the rest of the database functions can't perform computations on text fields.

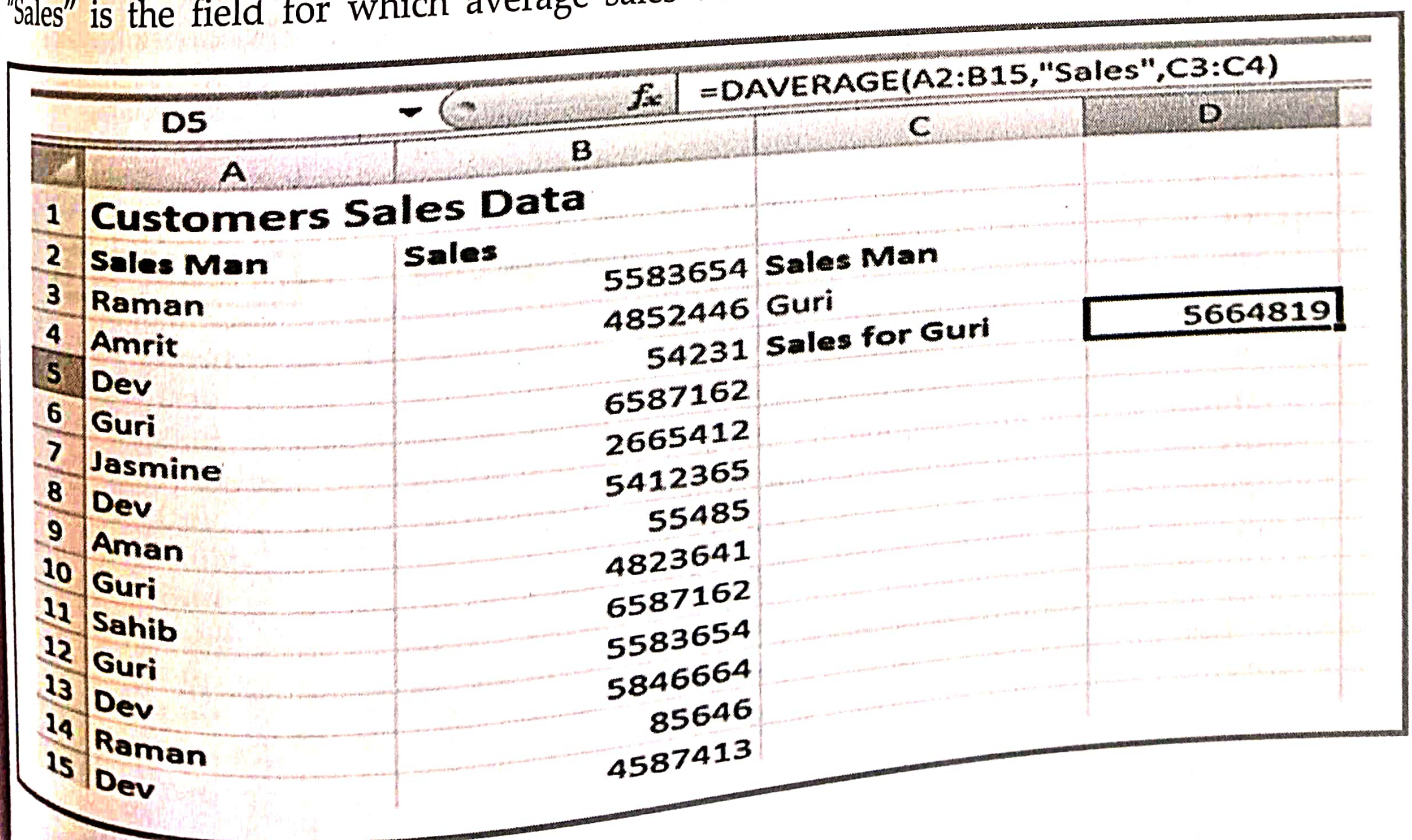
- **criteria is** the argument that specifies the address of the range that contains the criteria that you're using to determine which values are calculated. This range must include at least one field name that indicates the field whose values are to be evaluated and one cell with the values or expression to be used in the evaluation.

To use a database function successfully, you need to create a suitable range of cells that you can use for criteria. Excel expects this range of cells to be in a strict arrangement. Here are some rules you need to follow:

- Each condition requires two cells. One cell specifies the name of the field, and the other specifies the filter condition. For example, you can enter "Name" for the field name and "Aman" for the filter condition or value of the field.
- The cell with the filter condition must be directly under the cell with the field name. If you put them side by side, the database function gives you an error.
- You can apply more than one filter condition to the same field. Just place all the conditions in a neat stack under the cell with the field name.
- You can add conditions to multiple fields. However, you must place the filtering information for each field in a separate column.
- You can add as many filter conditions as you want

9.4.1. DAVERAGE FUNCTION

The DAVERAGE function lets you find the average, or mean, of a field for just the rows that match the criteria. For this example, you return to the customer sales database. Following figure shows a worksheet in which the average sale for salesmen "Guri" has been calculated by DAVERAGE function. In which cell A2:B15 is the range of database, "Sales" is the field for which average sales will be calculated and criteria is C3:C4.



Lets take another example in which average sale will be calculated for the sale > 6000000.

D5 =DAVERAGE(A2:B15,"Sales",C3:C4)

	A	B	C	D
1	Customers Sales Data			
2	Sales Man	Sales		
3	Raman	5583654	Sales	
4	Amrit	4852446	> 6000000	
5	Dev	54231	Average Sale	6587162
6	Guri	6587162		
7	Jasmine	2665412		
8	Dev	5412365		
9	Aman	55485		
10	Guri	4823641		
11	Sahib	6587162		
12	Guri	5583654		
13	Dev	5846664		
14	Raman	85646		
15	Dev	4587413		

9.4.2. DCOUNT FUNCTION

DCOUNT Counts the number of rows that meet the specified criteria. For example in customer sale data if we have to count how many sales have been made more than 6000000. Following figure shows the working of DCOUNT.

D5 =DCOUNT(A2:B15,"Sales",C3:C4)

	A	B	C	D
1	Customers Sales Data			
2	Sales Man	Sales		
3	Raman	5583654	Sales	
4	Amrit	4852446	> 6000000	
5	Dev	54231	Average Sale	2
6	Guri	6587162		
7	Jasmine	2665412		
8	Dev	5412365		
9	Aman	55485		
10	Guri	4823641		
11	Sahib	6587162		
12	Guri	5583654		
13	Dev	5846664		
14	Raman	85646		
15	Dev	4587413		

9.4.3. DMAX FUNCTION

DMAX returns the maximum value in rows that meet the specified criteria. For example to get maximum sale made by sales man function shows the in the following figure.

D5 =DMAX(A2:B15,"Sales",B3:B15)

	A	B	C	D
1	Customers Sales Data			
2	Sales Man	Sales		
3	Raman	5583654		
4	Amrit	4852446		
5	Dev	54231	Max Sale	6587777
6	Guri	6587777		
7	Jasmine	2665412		
8	Dev	5412365		
9	Aman	55485		
10	Guri	4823641		
11	Sahib	6587777		
12	Guri	5583654		
13	Dev	5846664		
14	Raman	85646		
15	Dev	4587413		

9.4.4. DMIN FUNCTION

DMIN returns the minimum value in rows that meet the specified criteria. For example to get minimum sale made by sales man function shows the in the following figure.

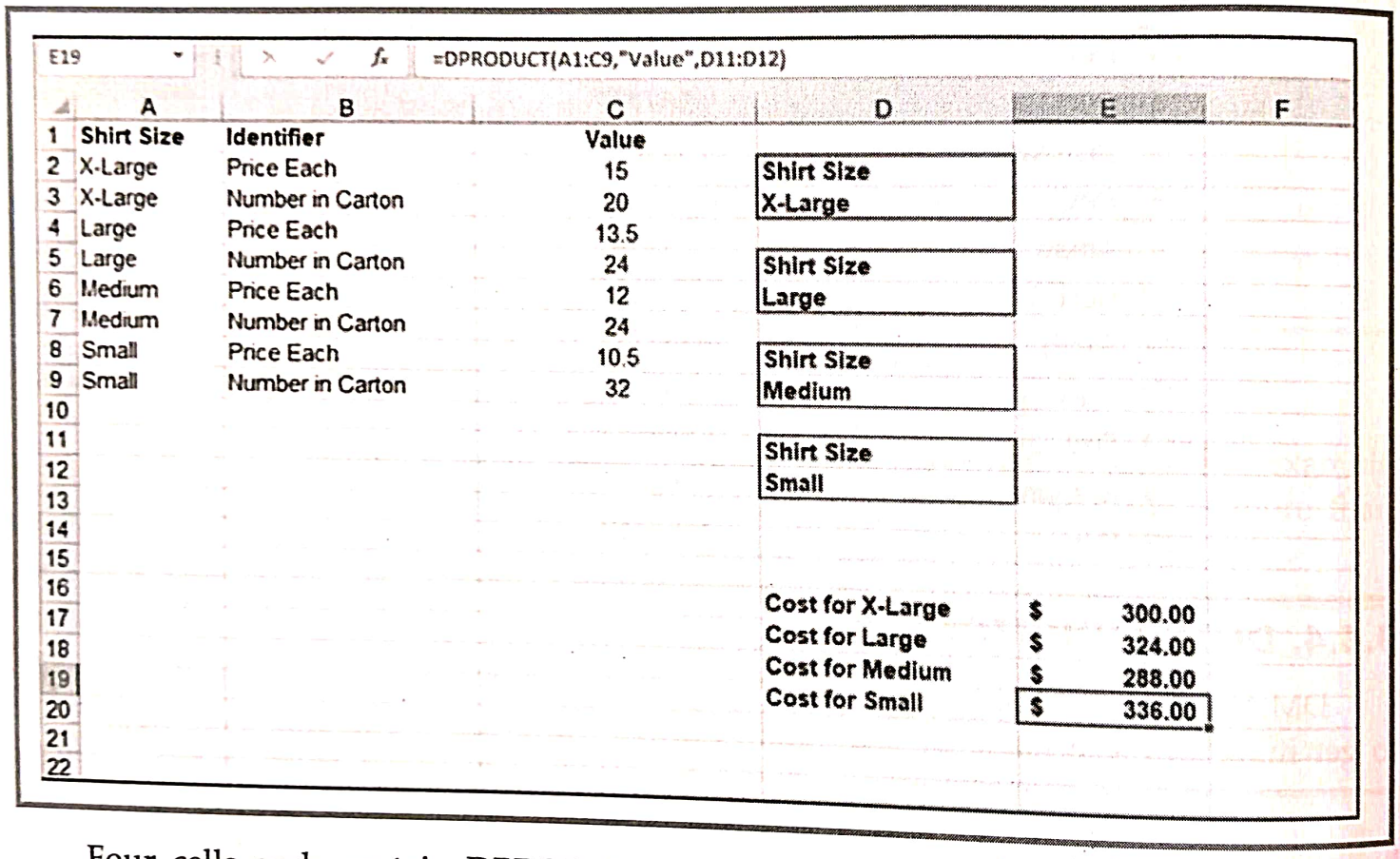
D5 =DMIN(A2:B15,"Sales",B3:B15)

	A	B	C	D
1	Customers Sales Data			
2	Sales Man	Sales		
3	Raman	5583654		
4	Amrit	4852446		
5	Dev	54231	Min Sale	54231
6	Guri	6587777		
7	Jasmine	2665412		
8	Dev	5412365		
9	Aman	55485		
10	Guri	4823641		
11	Sahib	6587777		
12	Guri	5583654		
13	Dev	5846664		
14	Raman	85646		
15	Dev	4587413		

9.4.5. DPRODUCT FUNCTION

DPRODUCT calculates the product produced by multiplying all values in rows that meet the specified criteria. **DPRODUCT** is less obviously useful than some of the other database functions, but it does have very specific invaluable niche uses.

Following figure shows a situation in which DPRODUCT is productive. The database area contains shirts. For each shirt size, there are two rows: the price per shirt and the number of shirts that are packed in a carton. The cost for a carton of shirts is, therefore, the product of the price per shirt times the number of shirts. There are four shirt sizes, each with its own price and carton count.



E19 =DPRODUCT(A1:C9,"Value",D11:D12)

	A	B	C	D	E	F
1	Shirt Size	Identifier	Value			
2	X-Large	Price Each	15	Shirt Size		
3	X-Large	Number in Carton	20	X-Large		
4	Large	Price Each	13.5			
5	Large	Number in Carton	24	Shirt Size		
6	Medium	Price Each	12	Large		
7	Medium	Number in Carton	24			
8	Small	Price Each	10.5	Shirt Size		
9	Small	Number in Carton	32	Medium		
10						
11				Shirt Size		
12				Small		
13						
14						
15						
16						
17				Cost for X-Large	$ 300.00	
18				Cost for Large	$ 324.00	
19				Cost for Medium	$ 288.00	
20				Cost for Small	$ 336.00	
21						
22						

Four cells each contain DPRODUCT, and within each cell, the particular criteria area is used. For example, cell E18 has this formula: =DPRODUCT(A1:C9, "Value", D8:D9). Four cells each contain DPRODUCT, and within each cell, the particular criteria area is used. For example, cell E18 has this formula: =DPRODUCT(A1:C9, "Value", D8:D9).

9.4.6. DGET FUNCTION

The DGET() function extracts the value of a single *field* in the *database* records that match the *criteria*. If there are no matching records, DGET() returns #VALUE!. If there's more than one matching record, DGET() returns #NUM!.

For example in the following figure to get the name of the product having No. 13, use the DGET method as show below:

E2 =DGET(A1:B10,E1,D1:D2)

	A	B	C	D	E	F	G	H	I
1	No.	Name		No.	Name				
2	12	carrots		13	salad				
3	13	salad							
4	14	bananas							
5	15	bread							
6	16	apples							
7	17	cabbage							
8	18	beef steak							
9	19	chicken							
10	20	cherries							
11									

9.4.7. DSUM FUNCTION

The DSUM function enables the Excel user to find a sum based on one or more criteria for a particular field in a database. DSUM is similar to SUMIF but more powerful in that it enables the Excel user to sum up values that meet multiple criteria in a dataset (versus a single criterion with the basic SUMIF function).

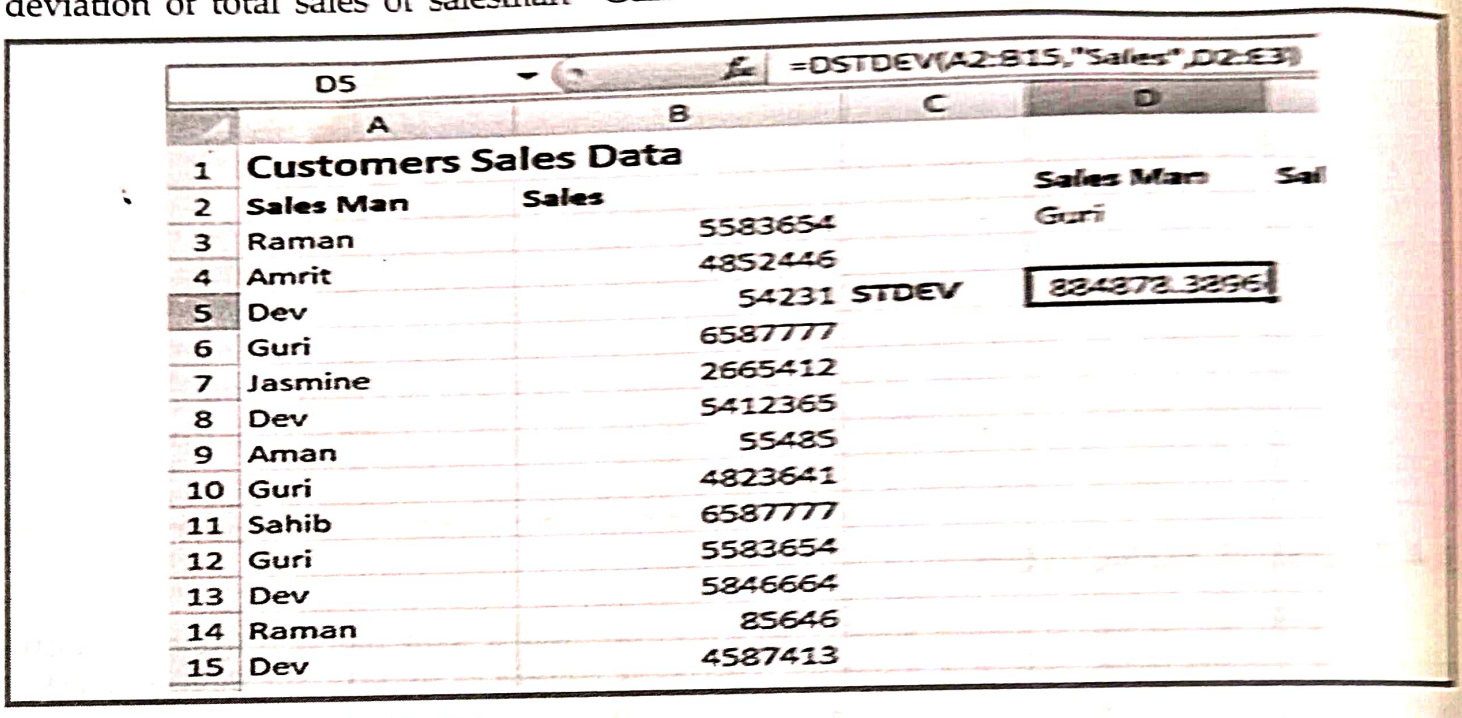
Following figure shows the total sale done by 'Guri' saleman.

D5 =DSUM(A2:B15,"Sales",D2:D3)

	A	B	C	D
1	Customers Sales Data			
2	Sales Man	Sales		
3	Raman	5583654		Sales Man
4	Amrit	4852446		Guri
5	Dev	54231	Total Sale	16995072
6	Guri	6587777		
7	Jasmine	2665412		
8	Dev	5412365		
9	Aman	55485		
10	Guri	4823641		
11	Sahib	6587777		
12	Guri	5583654		
13	Dev	5846664		
14	Raman	85646		
15	Dev	4587413		

9.4.8. DSTDEV FUNCTION

DSTDEV estimates the standard deviation of a population, based on a sample of selected entries from the database. The DSTDEV function estimates the deviation of a population based on entries in the list or database. Following figure calculate the standard deviation of total sales of salesman "Guri".

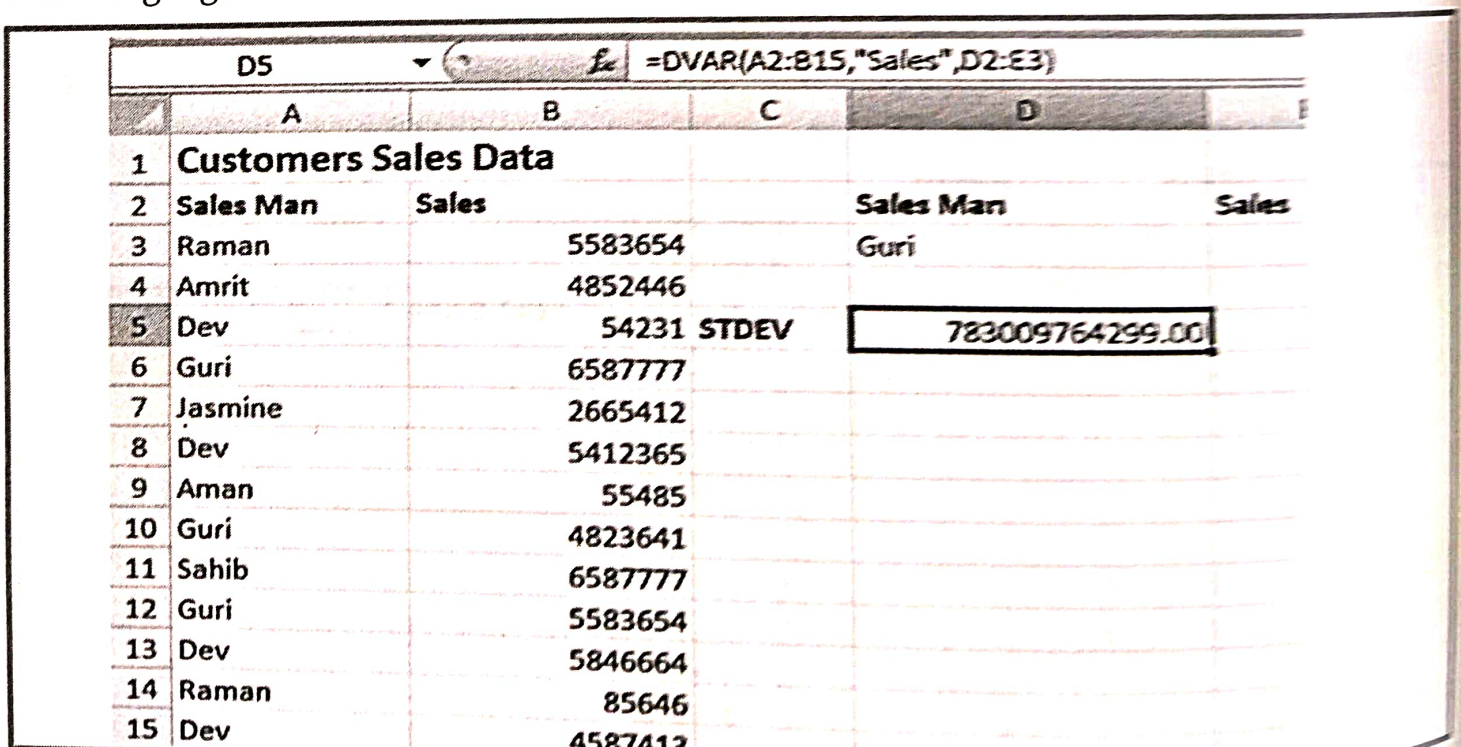


D5 =DSTDEV(A2:B15,"Sales",D2:E3)

	A	B	C	D	E
1	Customers Sales Data				
2	Sales Man	Sales		Sales Man	Sal
3	Raman	5583654		Guri	
4	Amrit	4852446			
5	Dev	54231	STDEV	884878.3896	
6	Guri	6587777			
7	Jasmine	2665412			
8	Dev	5412365			
9	Aman	55485			
10	Guri	4823641			
11	Sahib	6587777			
12	Guri	5583654			
13	Dev	5846664			
14	Raman	85646			
15	Dev	4587413			

9.4.9. DVAR FUNCTION

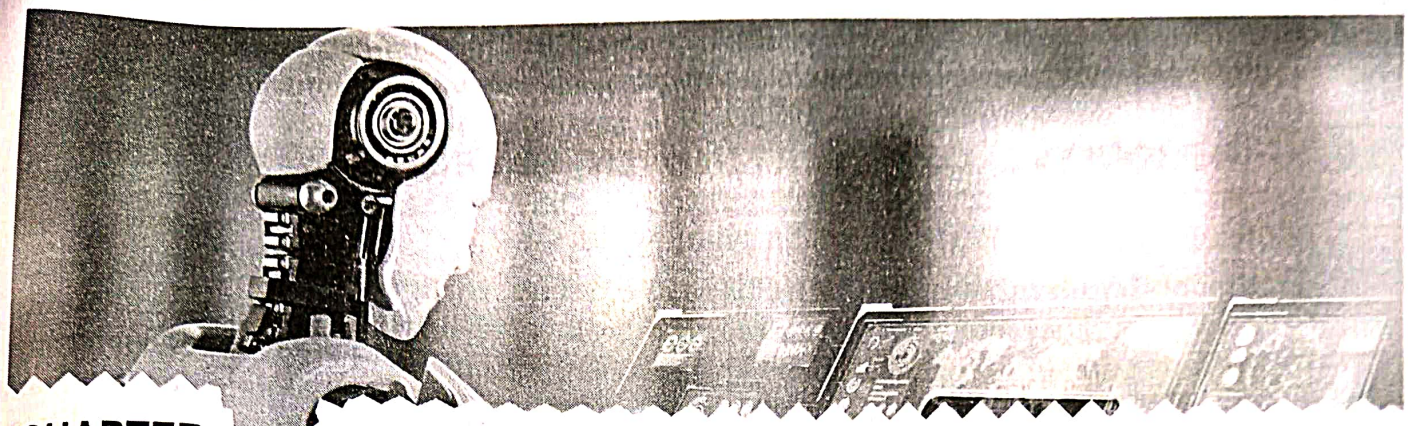
DVAR estimates the variance of a sample population, based on the values that meet the specified criteria. The DVAR function provides the variance of a sample population. Following figure calculate the variance of Guri salesman of the customer sales data.



D5 =DVAR(A2:B15,"Sales",D2:E3)

	A	B	C	D	E
1	Customers Sales Data				
2	Sales Man	Sales		Sales Man	Sales
3	Raman	5583654		Guri	
4	Amrit	4852446			
5	Dev	54231	STDEV	783009764299.00	
6	Guri	6587777			
7	Jasmine	2665412			
8	Dev	5412365			
9	Aman	55485			
10	Guri	4823641			
11	Sahib	6587777			
12	Guri	5583654			
13	Dev	5846664			
14	Raman	85646			
15	Dev	4587413			

☆☆☆☆☆



CHAPTER

10 CHARTS AND PIVOT TABLE

CHAPTER AT A GLANCE

10.1 Introduction
10.2 Objective of Charting in Excel
10.3 Advantages of Charting
10.4 Limitation of Charting
10.5 Different Ways to Create Chart
10.6 Parts of Chart
10.7 Common Chart Types in Excel
10.8 Creating Embedded Chart
10.9 Formatting Chart
10.10 Changing Chart Type
10.11 Adding Title, Legend and Gridlines
10.12 Printing Chart
10.13 Introduction to Pivot Table
10.14 Importance of Pivot Tables
10.15 Advantages of Pivot Tables
10.16 Applications of Pivot Tables
10.17 Building Pivot Table
10.18 Different Levels to Prepare the Pivot Table
10.19 Designing Pivot Table
10.20 Formatting Pivot Table
10.21 Grouping Fileds in Pivot Table
10.22 Slicing Data in Pivot Table
10.23 Filtering Data in Pivot Table
10.24 Sorting Data in Pivot Table

10.1 INTRODUCTION

Charts depict data visually, so you can quickly spot trends. They're a fabulous way to help you find the meaning hidden in large amounts of data. You can create many different types of charts in Excel, including pie charts that present polling results, line charts that plot rising or declining assets over time, and three-dimensional area charts that show relationships between environmental conditions in a scientific experiment.

A chart is essentially an object that Excel creates upon request. This object consists of one or more data series, displayed graphically. The appearance of the data series depends on the selected chart type. *The data for each series is stored in a separate row or column. Each*

point on the line is determined by the value in a single cell and is represented by a marker. You can distinguish each of the lines by its thickness, line style, color, or data markers.

10.2 OBJECTIVE OF CHARTING IN EXCEL

Following are the objectives for the charting in the Excel.

1. **Visual Representation of Data:** Charts in Excel serve primarily to visually represent data sets. They convert numerical data from spreadsheets into graphical forms such as bars, lines, pie slices, etc. This visual representation helps in understanding the data at a glance, which is often more intuitive and easier to interpret than looking at raw numbers alone. By presenting data graphically, charts facilitate quick insights into trends, patterns, and relationships within the data.
2. **Facilitating Data Analysis**: One of the key objectives of using charts in Excel is to facilitate data analysis. Charts allow users to analyze large datasets more efficiently by visually highlighting important information such as highs, lows, and trends. For example, a line chart can show how a particular metric has changed over time, enabling analysts to identify patterns or anomalies that may require further investigation. This analytical capability helps in making informed decisions based on data-driven insights.
3. **Comparison and Trends Identification:** Charts enable easy comparison between different data sets or categories. Whether comparing sales figures between different quarters or examining the distribution of expenses across departments, charts provide a clear visual depiction of comparisons. Moreover, they help in identifying trends over time or across different variables. Trends such as growth rates, seasonal variations, or shifts in consumer preferences become apparent when data is plotted on charts, aiding in strategic planning and forecasting.
4. **Effective Communication:** Charts play a crucial role in communicating findings and insights to stakeholders and colleagues. Instead of presenting complex tables or lengthy reports, charts offer a concise and visually appealing way to convey information. They enhance communication by making data more accessible and understandable to a wider audience, including those who may not be familiar with the specifics of the data being presented. This ability to communicate effectively through charts fosters better decision-making and collaboration within teams.
5. **Supporting Decision-Making Processes**: Another important objective of charts in Excel is to support decision-making processes. By presenting data visually, charts help decision-makers quickly grasp the implications of different choices or scenarios. For instance, a bar chart comparing revenue from different product lines can assist in identifying which products are performing well and which ones may need attention. This clarity and transparency in data presentation empower decision-makers to make informed and timely decisions based on credible data insights.

In conclusion, charts in Excel serve multiple objectives ranging from enhancing data visualization and analysis to facilitating effective communication and supporting decision-

making processes. They transform raw data into meaningful visuals that not only simplify complex information but also empower users to derive actionable insights and drive informed decisions across various domains of business and analysis.

10.3 ADVANTAGES OF CHARTING

1. **Graphical Representation:** Charts in Excel provide a clear and visual representation of data, making complex numerical data easier to interpret and understand at a glance. Visual elements such as bars, lines, and pie slices help in identifying trends, patterns, and outliers within the data quickly and intuitively.
2. **Supports Data Analysis:** Excel charts facilitate effective data analysis by summarizing large datasets into graphical forms. They allow users to analyze trends over time, compare different categories or variables, and identify correlations or discrepancies more efficiently than analyzing raw data in tables.
3. **Enhances Communication:** Charts in Excel enhance communication of data insights to stakeholders, colleagues, or clients. They provide a concise and visually appealing way to present findings, making it easier for non-technical audiences to grasp complex information and make informed decisions based on data-driven insights.
4. **Enhance Decision-Making:** By presenting data visually, Excel charts support decision-making processes. Decision-makers can quickly assess the performance of various metrics, identify areas needing attention, and formulate strategies based on trends and patterns highlighted in the charts. This improves the speed and accuracy of decision-making within organizations.
5. **Versatility and Customization:** Excel offers a wide range of chart types (e.g., column, line, pie, scatter, etc.) and customization options. Users can customize colors, labels, axes, and other elements to suit specific needs and preferences. This versatility allows for the creation of charts that best represent the data and convey the intended message effectively.

10.4 LIMITATION OF CHARTING

1. **Potential for Misinterpretation:** While charts simplify data presentation, there is a risk of misinterpretation if they are not designed or labeled clearly. Incorrect scaling, misleading visuals, or inadequate data labeling can lead to misunderstandings and incorrect conclusions drawn from the data presented.
2. **Complexity with Large Datasets:** Excel charts may struggle with large datasets or data that requires complex visual representation. Performance issues can arise when attempting to plot extensive datasets or when trying to visualize data with multiple variables or dimensions, potentially leading to slower processing times or charts that are difficult to interpret.
3. **Overemphasis on Visual Appeal:** There is a tendency to prioritize visual appeal over data accuracy or relevance when creating charts in Excel. Users may focus

on making charts aesthetically pleasing rather than ensuring they accurately represent the underlying data or provide meaningful insights, which can undermine the credibility and usefulness of the charts.

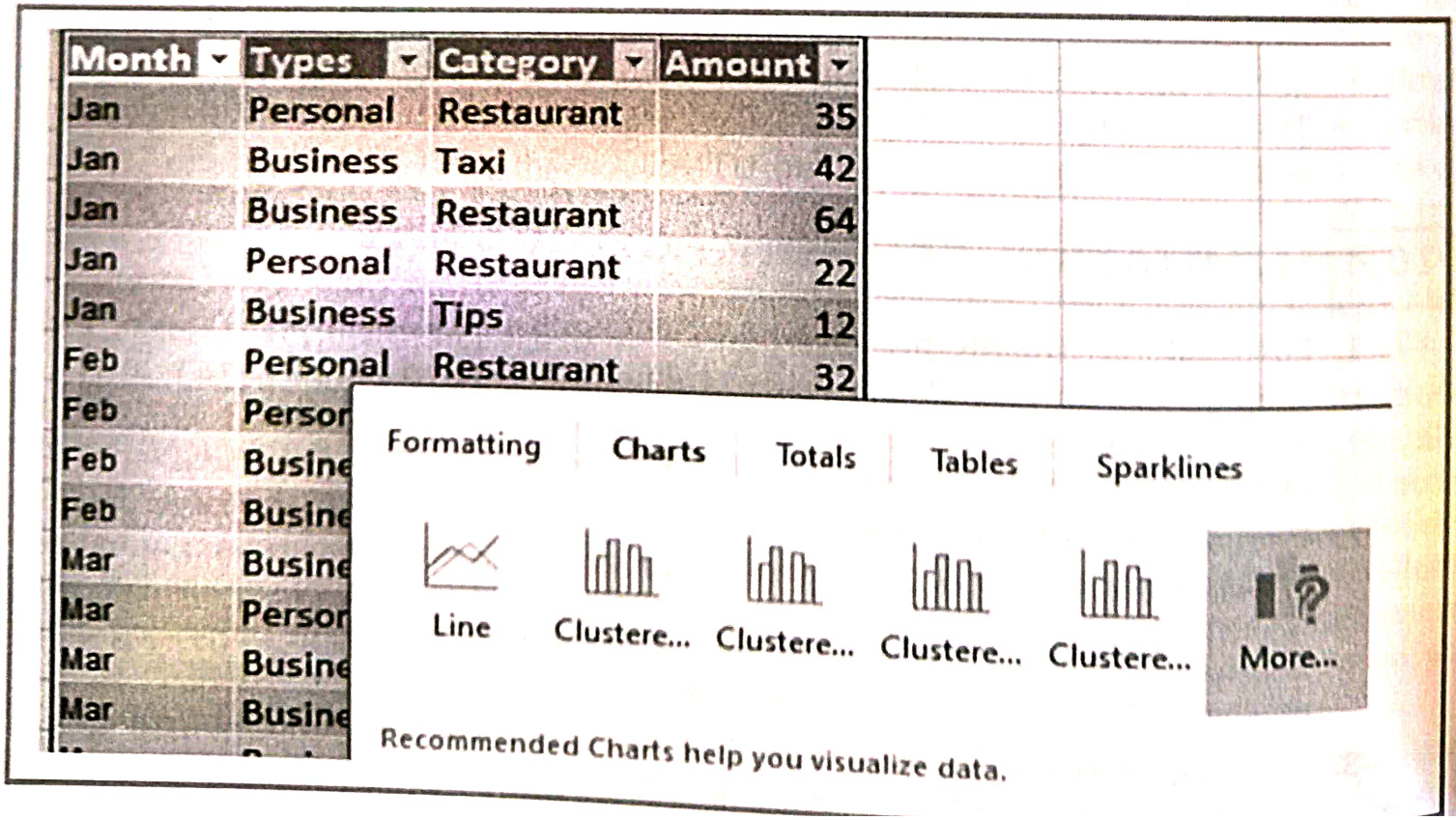
4. **Limited Interactivity:** Excel charts, particularly static ones in printed or static digital formats, lack interactive features that allow users to explore data dynamically. Unlike more advanced data visualization tools, Excel charts do not always support interactive filtering, drill-down capabilities, or real-time updates, limiting their usefulness in dynamic data analysis contexts.
5. **Dependency on Software and Skillset:** Effective utilization of Excel charts depends on users' proficiency with Excel and its charting features. Creating advanced or custom charts may require specific skills or knowledge of Excel functions, which not all users may possess. Moreover, charts created in Excel may not always be easily transferable or compatible with other software platforms or data visualization tools.

10.5 DIFFERENT WAYS TO CREATE CHART

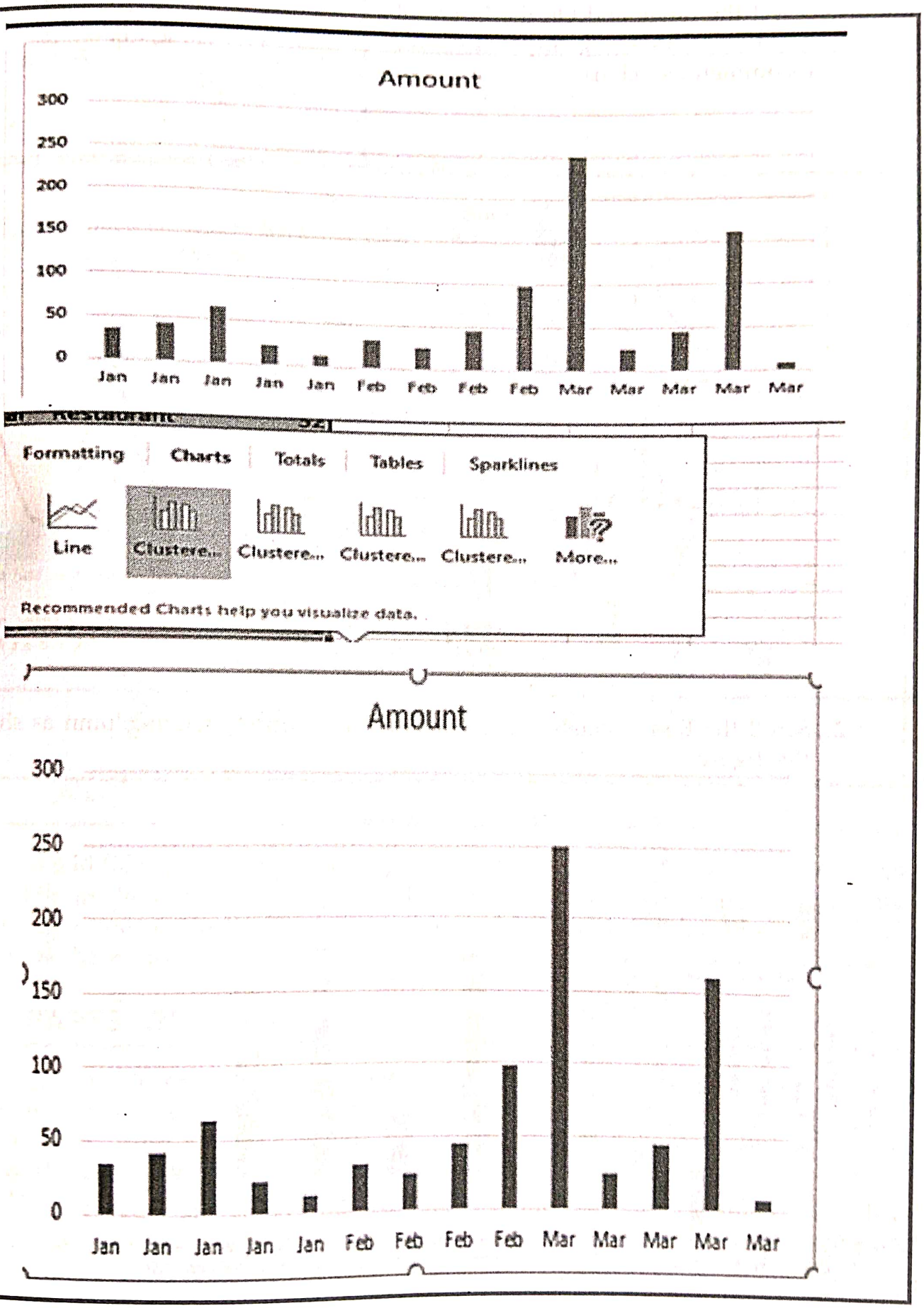
10.5.1. USING QUICK ANALYSIS TOOL

You can create charts quickly by using the Quick Analysis Lens, which displays recommended charts to summarize your data. When you select the entire data range you want to chart, clicking the Quick Analysis action button lets you display the types of charts Excel recommends. Following steps shows the to create the chart using quick analysis tool.

Select the data you want to appear in your chart. Click the *Quick Analysis* action button.

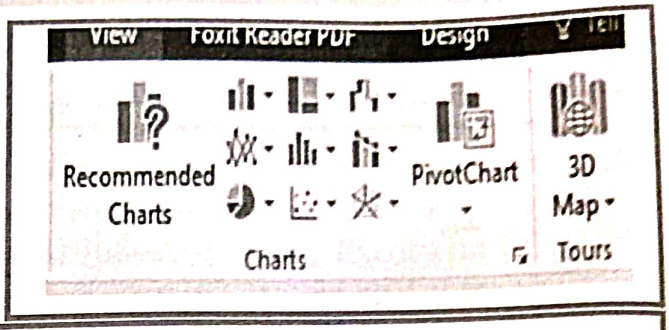


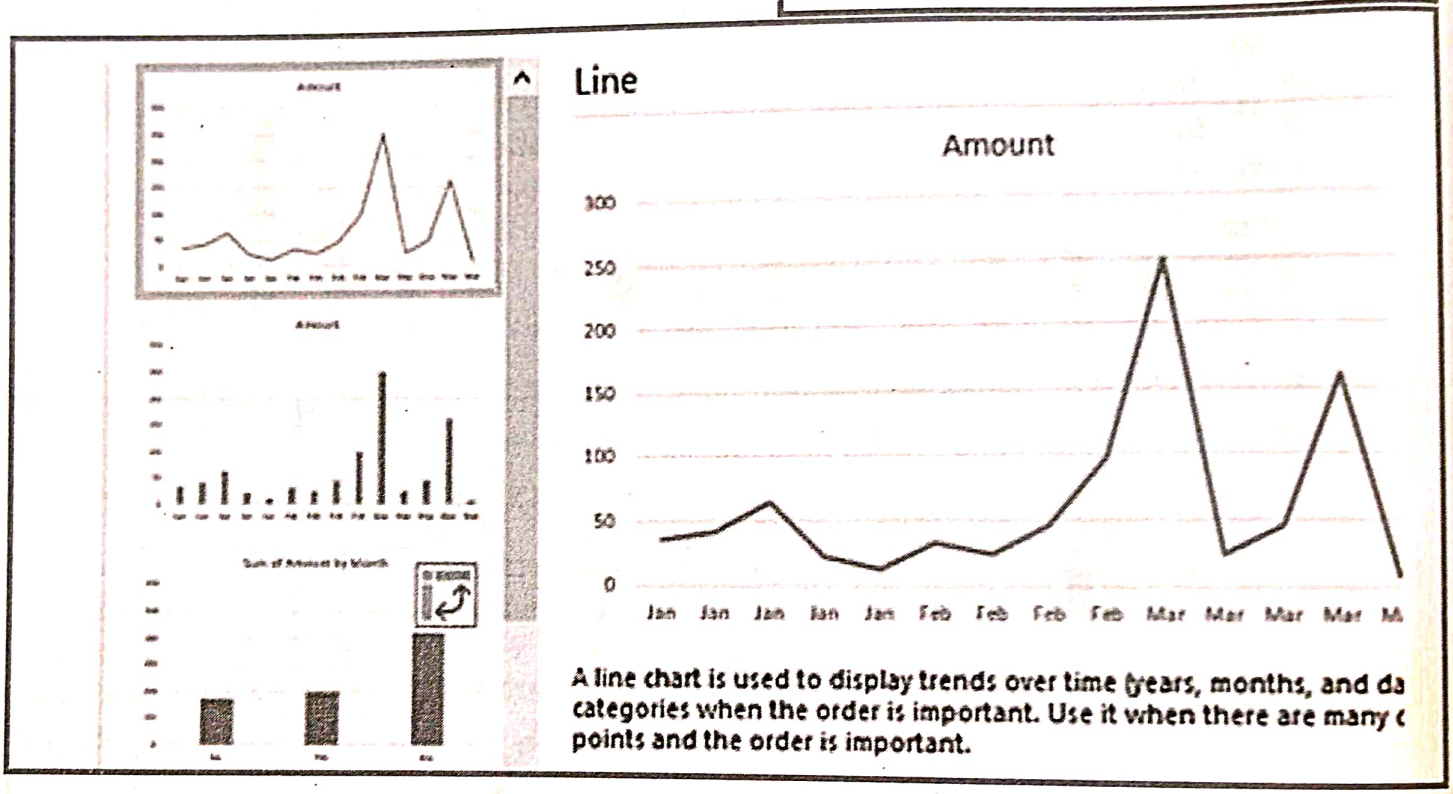
In the gallery that appears, click the *Charts* tab. Click the chart type you want to create.



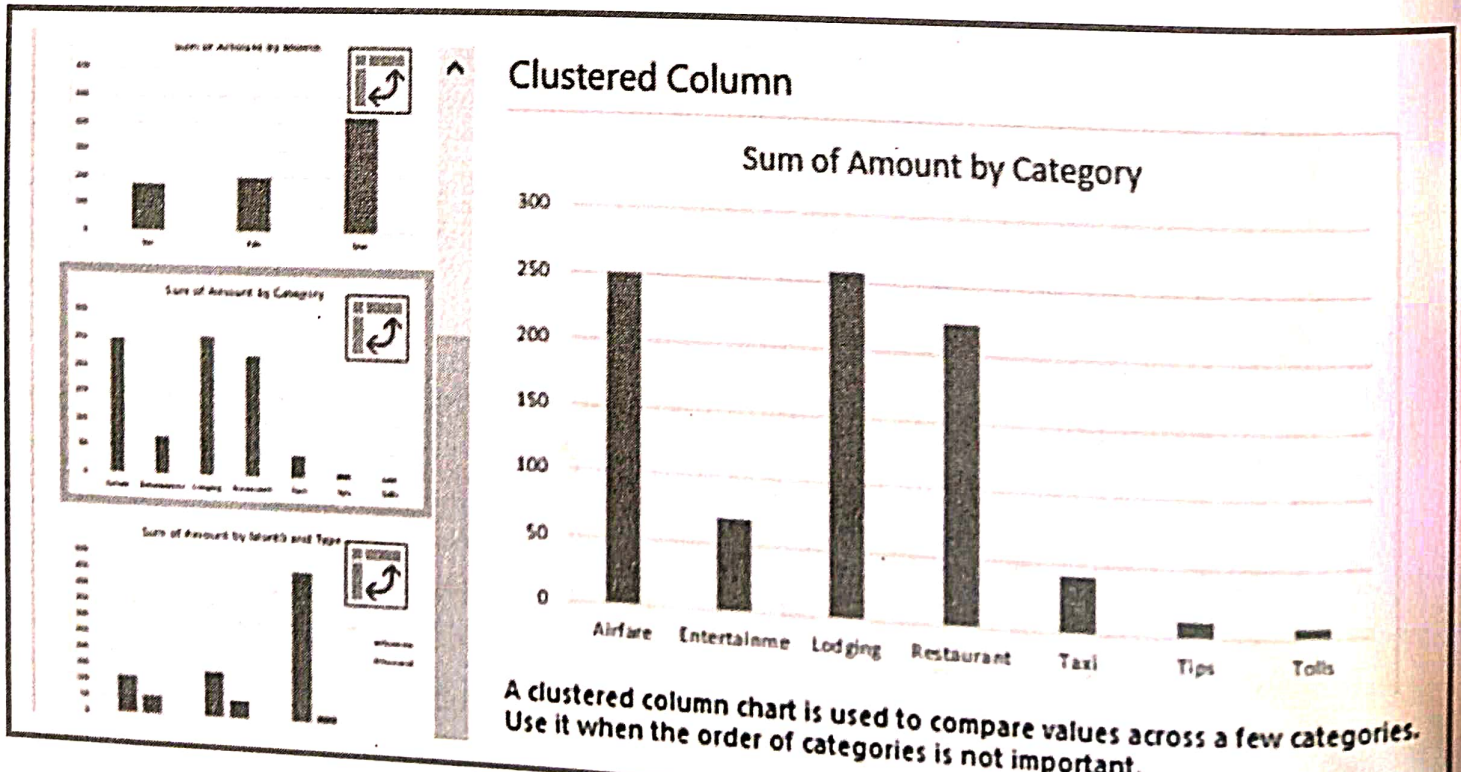
10.5.2. INSERTING CHART

1. Select the range and On the Insert tab, in the Charts group click on recommended charts.

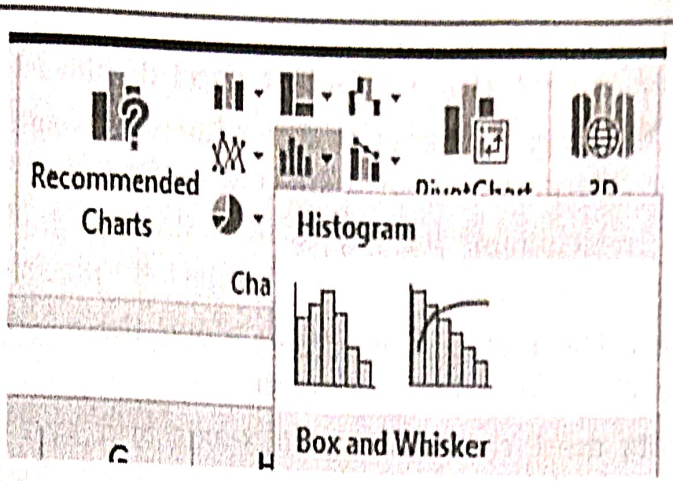


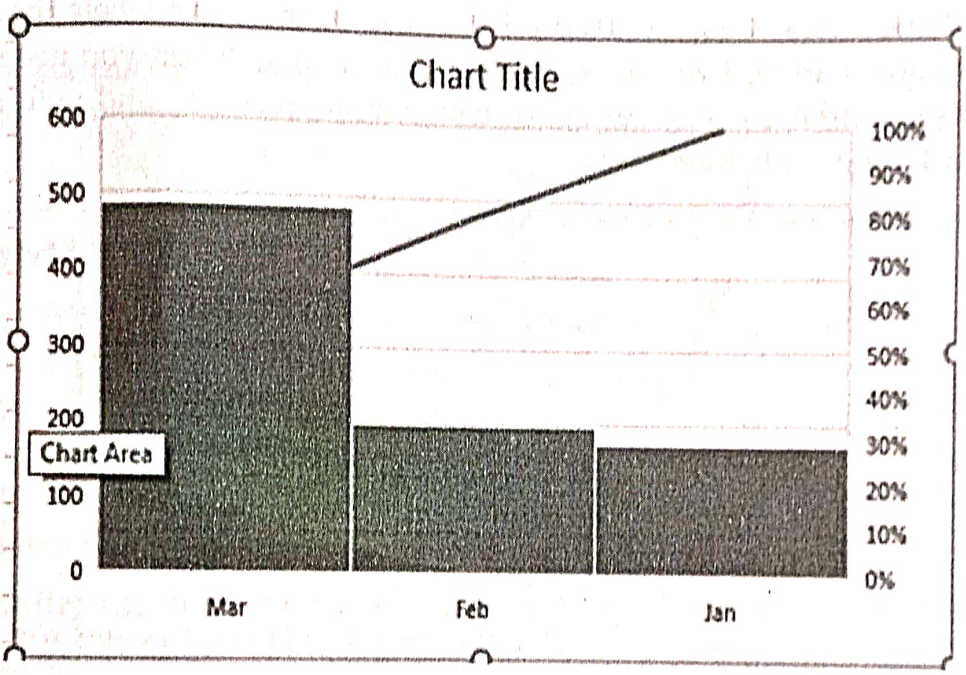


2. Scroll the Recommended chart column and select clustered column as shown in the figure.



3. Alternatively, you can choose chart type directly from the chart group as shown below





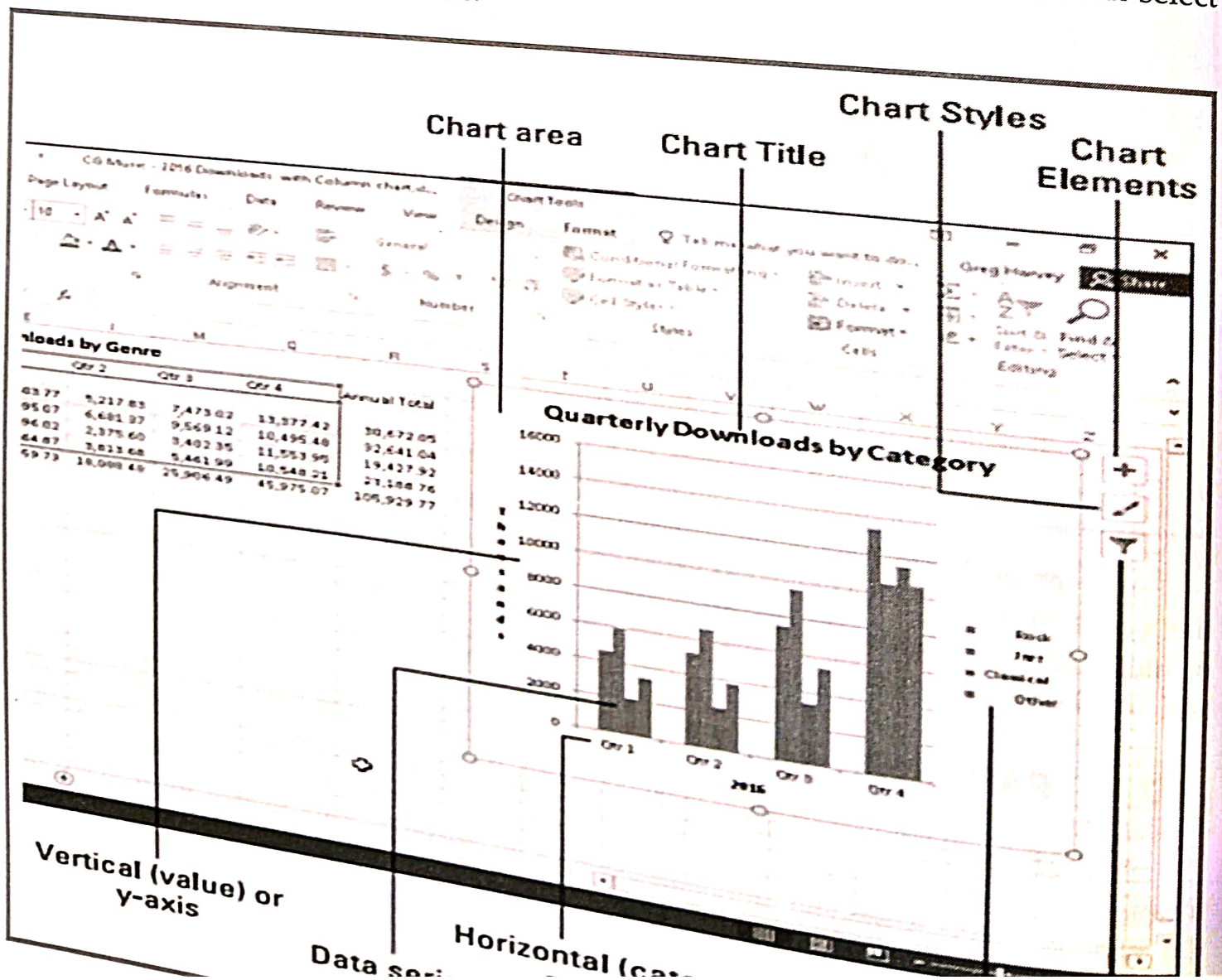
You can add titles to a chart to identify the chart's subject. By default, Excel designates the x-axis title as the categories that you include in your data, and they-axis as the unit of measure that you use in the chart. If you create a 3-D chart, Excel may allow you to add a title to the z-axis.

10.6 PARTS OF CHART

A chart is made of different elements, each of which contributes to the overall meaning and appearance of the chart. You can customize each element as indicated, using the various tasks. After you understands how to create a chart to you must know the different parts of a chart.

- **Data Marker:** The individual value represented by one number from the worksheet and charted as one element in the chart. A collection of data markers makes up a data series.

- **Chart Title:** You can use chart titles to label the various parts of your chart, including the name of the chart and the x- and y-axis.
- **Chart Area:** The area on which Excel draws your chart. You can change the background color as well as the lines that Excel displays to represent your data.
- **Data Series:** A collection of one set of data values. For example, a data series may include sales for one particular year in your company.
- **Legend:** The legend contains the key to the values in the chart. You can choose whether to display the legend, as well as select its placement.
- **Category Names:** The names that appear along the x-, or category axis. Excel pulls these categories from the worksheet.
- **Axes:** Excel plots most charts using an x-axis and ay-axis — also called the horizontal and vertical axes. Excel also has chart types that include a third axis, called the z-axis. You can choose how Excel formats and labels the various axes. If you are charting two points, two axes are enough. When you need to plot three points or variables — for example, temperature, pressure, and volume then select a chart type with three axes.



10.7 COMMON CHART TYPES IN EXCEL

10.7.1 BAR CHART

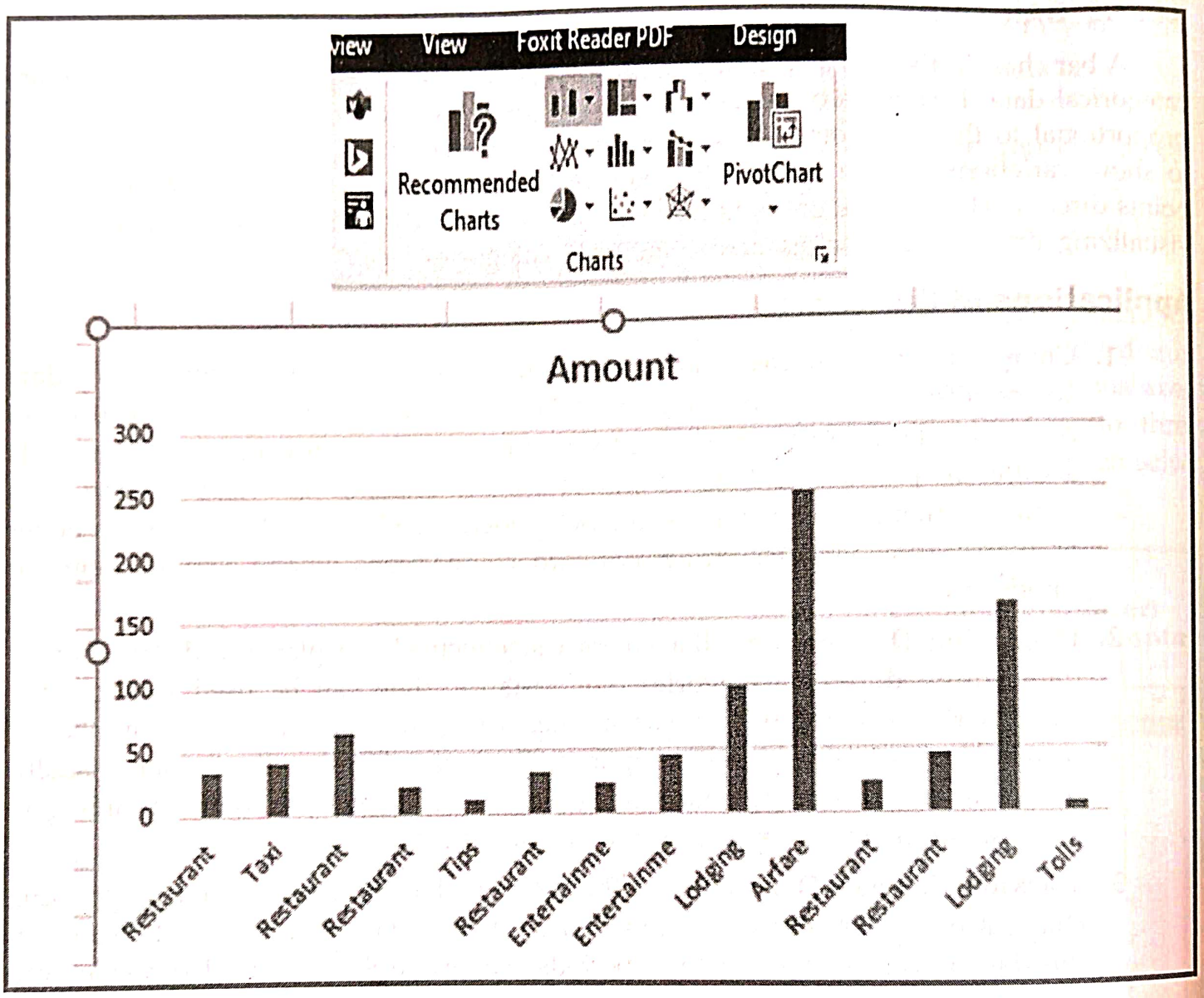
A bar chart in Excel is a type of graphical representation used to display and compare categorical data. It consists of rectangular bars, where the length or height of each bar is proportional to the value or frequency of the data it represents. Bar charts are designed to show variations in size among different categories, making it easy to compare data points directly. They can be oriented horizontally or vertically and are especially useful for visualizing discrete data points across multiple categories.

Applications of Bar Chart:

1. **Comparing Data Across Categories:** Bar charts are ideal for comparing data across different categories or groups. For example, a bar chart can be used to compare sales figures of various products within a company. Each bar represents a different product, and the length of the bar indicates the sales volume. This visual comparison allows users to quickly identify which products perform better or worse and facilitates straightforward comparisons among various items or categories.
2. **Displaying Distributions:** Bar charts can effectively display the distribution of data within different categories. For instance, they can be used to show the frequency of occurrences for various items in a survey, such as the number of respondents selecting different options. By presenting the distribution visually, users can easily observe which categories are most or least common, aiding in understanding the overall distribution pattern.
3. **Tracking Changes Over Time:** Although line charts are often used for tracking changes over time, bar charts can also be used to visualize temporal changes if the data is categorized by time periods. For example, a vertical bar chart can show monthly sales figures across several months. Each bar represents a month, allowing users to observe how values fluctuate over time and identify trends or patterns in the data.
4. **Evaluating Performance Metrics:** In performance analysis, bar charts can be employed to evaluate various metrics, such as employee performance, project progress, or departmental achievements. For instance, a bar chart might display individual employee performance ratings or project completion percentages. This visual format aids in assessing performance across different individuals or projects and supports data-driven decision-making and performance evaluation.
5. **Presenting Statistical Data:** Bar charts are often used to present statistical data, such as the mean, median, and mode of various datasets. By visualizing these statistical measures, users can quickly grasp the central tendency and variability within the data. For example, a bar chart might show the average test scores across different classes, allowing educators to compare performance levels easily.

Adding Bar Chart

Select the range and On the Insert tab, in the Charts group click on Bar charts.



10.7.2 PIE CHART

A pie chart in Excel is a type of circular graph that is divided into slices to illustrate numerical proportions. Each slice represents a category's contribution to the total, with the size of each slice being proportional to the value it represents relative to the whole pie. Pie charts are used to show the relative sizes of parts to a whole, making them particularly useful for illustrating percentage distributions among various categories. This visualization helps users quickly understand the proportion of each category within a dataset.

Applications of Pie Charts:

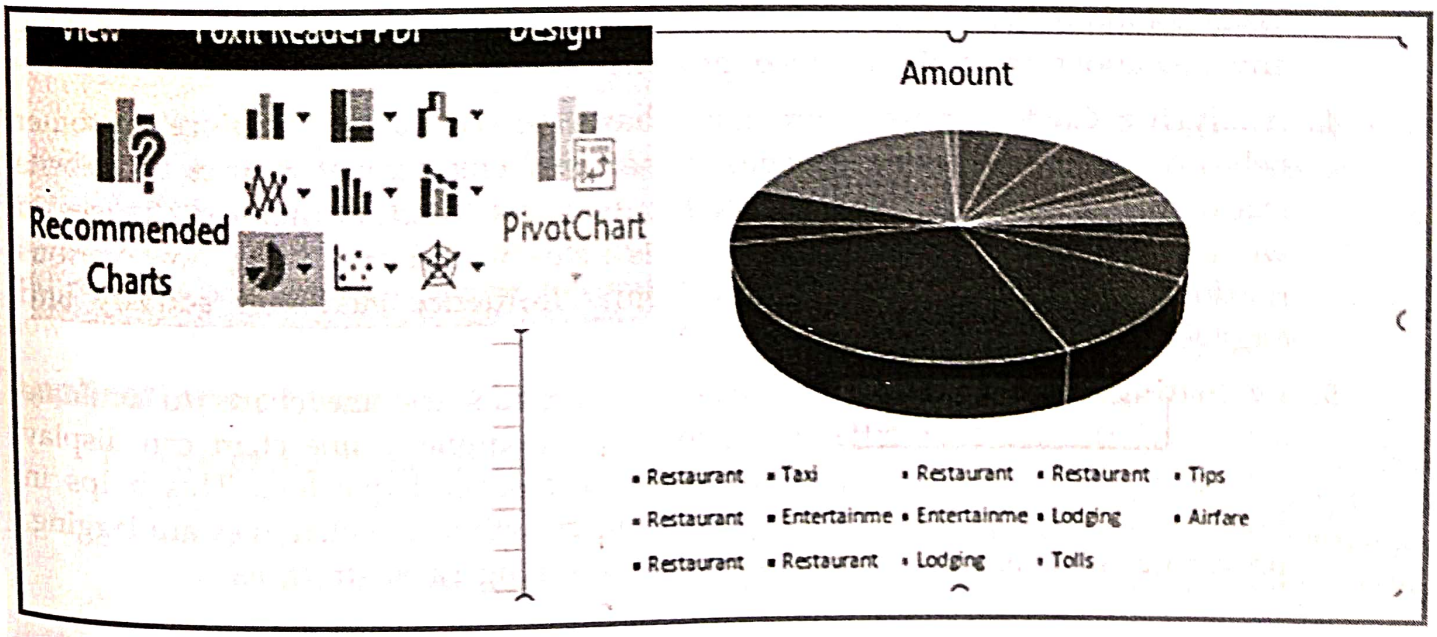
1. **Displaying Proportional Data:** Pie charts are ideal for displaying data as a proportion of a whole. For example, if a company wants to show the market share of different products, a pie chart can illustrate each product's contribution to the total market share. Each slice of the pie represents a product, with the size of the slice indicating its market share percentage. This visual representation

helps in understanding how different categories compare to each other within the total.

2. **Showing Market Share Distribution:** Businesses often use pie charts to show the distribution of market share among competitors. For example, a pie chart could illustrate the market share of different companies within an industry. Each slice represents a company's share of the market, making it easy to compare the relative size of each competitor's presence in the market.
3. **Highlighting Product Sales Proportions:** Pie charts can be used to highlight the sales proportions of various products within a company. For instance, if a retailer wants to showcase the contribution of different product lines to total sales, a pie chart can depict each product line's share of overall sales. This visual representation helps in understanding which products are performing well and which are under performing.
4. **Demonstrating Demographic Data:** Pie charts are useful for presenting demographic data, such as age distribution, gender ratios, or geographic regions. For example, a pie chart could show the distribution of employees across different age groups within a company. This type of visualization provides a clear and immediate understanding of demographic composition, which can be valuable for human resources planning and strategy.
5. **Illustrating Project Task Distribution:** In project management, pie charts can illustrate the distribution of tasks or resources among different project components or team members. For example, a pie chart could show the percentage of total project time allocated to various tasks or phases. This helps in visualizing how resources are divided and ensures that all aspects of the project are accounted for in terms of effort and time.

Adding Pie Chart

Select the range and On the Insert tab, in the Charts group click on Pie charts.



10.7.3 LINE CHART

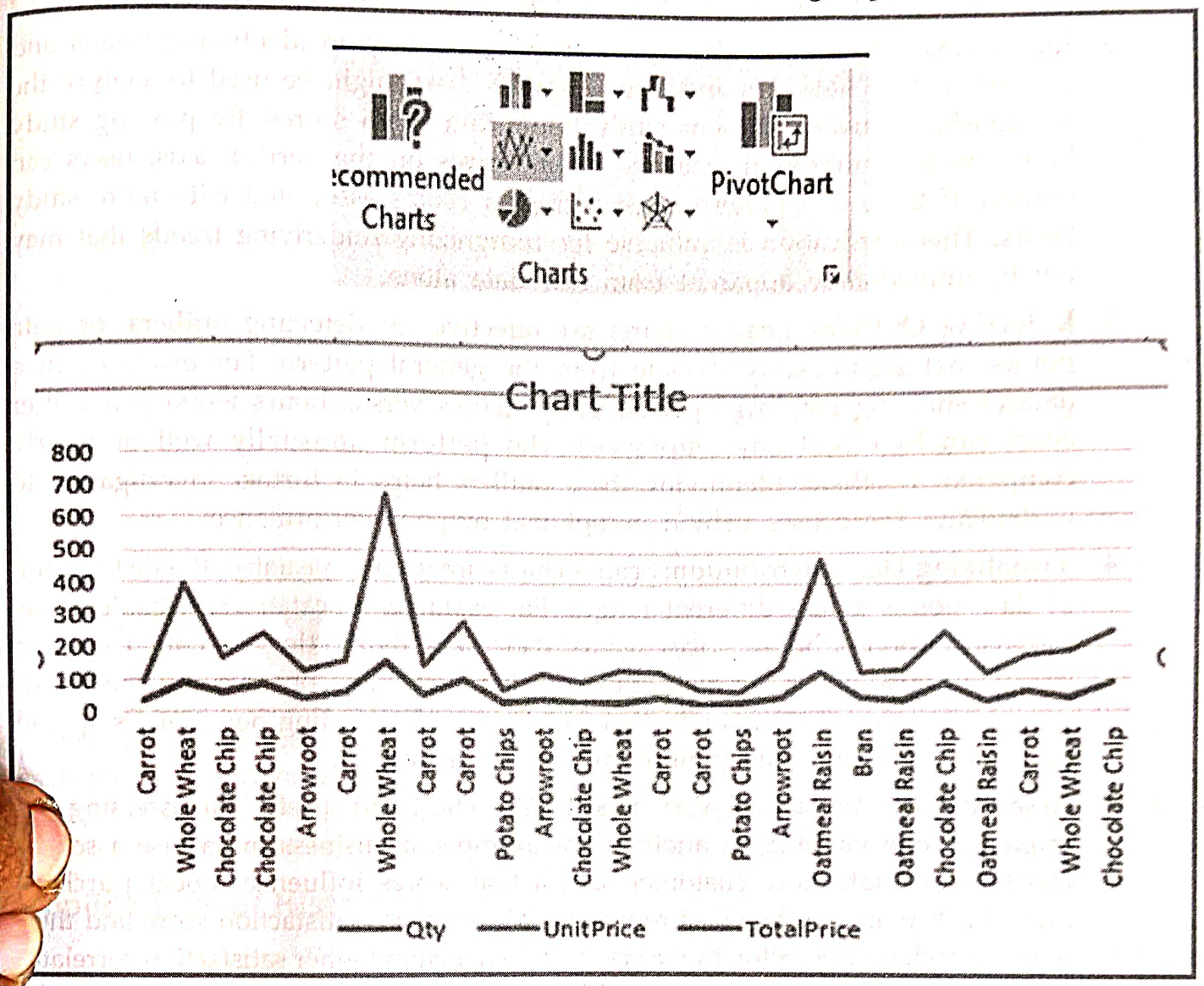
A line chart in Excel is a type of graph that uses lines to connect data points plotted on a Cartesian plane, typically with time on the horizontal axis and values on the vertical axis. This type of chart is ideal for visualizing trends and patterns over time, as it highlights changes in data values at various intervals. Each line in a line chart represents a series of data, allowing users to track fluctuations, compare multiple data series, and observe how values evolve.

Applications of Line Charts:

1. **Tracking Trends Over Time**: Line charts are particularly effective for tracking changes and trends over time. For example, a business might use a line chart to display monthly sales figures over a year. By plotting sales data on the vertical axis and time (months) on the horizontal axis, the line chart provides a clear view of how sales figures have increased or decreased over the year. This visualization helps in identifying seasonal patterns, growth trends, or declining performance.
2. **Comparing Multiple Data Series**: Line charts are useful for comparing multiple data series within the same chart. For instance, a line chart can be used to compare the performance of different sales regions over the same period. Each region's sales data can be represented by a separate line, allowing for direct comparison of performance trends between regions. This application is valuable for assessing relative performance and making strategic decisions based on comparative analysis.
3. **Monitoring Project Progress**: Line charts can be employed to monitor progress on projects by tracking key performance indicators (KPIs) or milestones over time. For example, a project manager might use a line chart to track the completion percentage of project tasks against planned deadlines. This visualization helps in assessing whether the project is on schedule and identifying any deviations from the planned progress.
4. **Analyzing Customer Behavior**: Line charts can be used to analyze customer behavior, such as tracking changes in customer engagement metrics or website traffic. For instance, a company might use a line chart to monitor changes in website visits over several months. This helps in understanding how various marketing campaigns or external factors influence customer activity and engagement.
5. **Evaluating Sales Performance**: Sales teams often use line charts to evaluate sales performance over different periods. For example, a line chart can display sales figures for different product lines over several quarters. This helps in identifying which products are experiencing growth and which ones are lagging, providing insights into sales trends and informing sales strategies.

Adding Line Chart

Select the range and On the Insert tab, in the Charts group click on Line charts.



10.7.4 SCATTER CHART

A scatter chart in Excel is a type of graph that displays data points as individual markers on a Cartesian plane, where each point represents the intersection of values for two different variables. The horizontal axis typically represents one variable, while the vertical axis represents another. Scatter charts are used to identify relationships, correlations, and patterns between two quantitative variables. They are particularly useful for examining how one variable affects another and for revealing clusters or trends within data.

Applications of Scatter Charts:

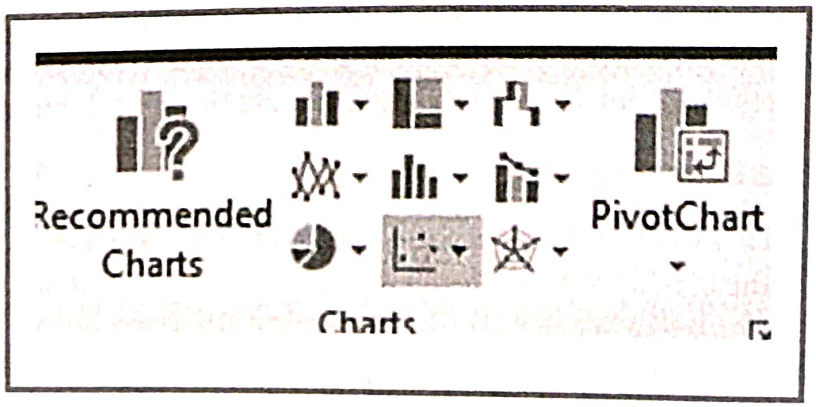
1. **Analyzing Relationships Between Variables**: Scatter charts are ideal for analyzing relationships between two quantitative variables. For example, a researcher might use a scatter chart to examine the correlation between advertising expenditure and sales revenue. Each point on the scatter chart represents a pair

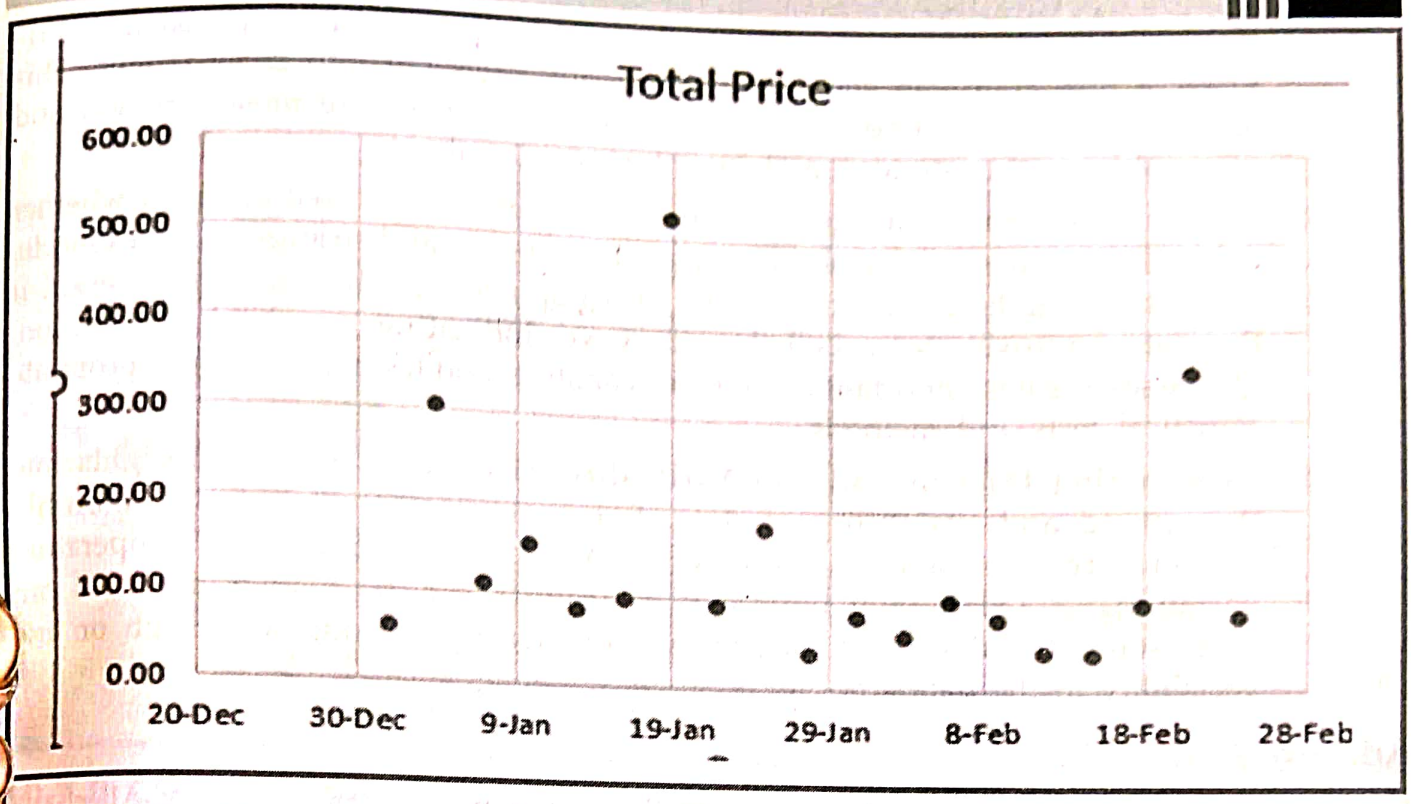
of values (advertising spend and sales revenue). By observing the distribution of data points, users can determine if there is a positive, negative, or no correlation between the variables, facilitating deeper insights into their relationship.

2. **Identifying Trends and Patterns:** Scatter charts help in identifying trends and patterns within data. For instance, a scatter chart might be used to analyze the relationship between student study hours and exam scores. By plotting study hours on the horizontal axis and exam scores on the vertical axis, users can observe if there is a pattern, such as higher scores associated with more study hours. This application is valuable for recognizing underlying trends that may not be immediately apparent from raw data alone.
3. **Detecting Outliers:** Scatter charts are effective for detecting outliers, or data points that significantly deviate from the general pattern. For example, in a dataset showing employee performance metrics versus hours worked, a scatter chart can highlight any employees who perform unusually well or poorly compared to others. Identifying these outliers helps in further investigation to understand the causes behind exceptional or poor performance.
4. **Visualizing Data Distribution:** Scatter charts are used to visualize the distribution of data points across different ranges. For example, a scatter chart could show the distribution of house prices against their sizes. By plotting the data, users can see how house prices are distributed relative to size and observe any clusters or gaps in the data. This visualization assists in understanding how data is spread and helps in identifying patterns within the dataset.
5. **Assessing the Impact of Variables:** Scatter charts are useful for assessing the impact of one variable on another. For example, a business might use a scatter chart to evaluate how customer satisfaction scores influence repeat purchase rates. Each point on the chart represents a customer's satisfaction score and their repeat purchase rate, allowing users to assess whether higher satisfaction correlates with more frequent purchases. This application helps in understanding the effectiveness of customer satisfaction initiatives.

Adding Scatter Chart

Select the range and On the Insert tab, in the Charts group click on Scatter charts.





10.7.5 HISTOGRAM CHART

A histogram chart in Excel is a type of graph used to represent the distribution of numerical data across different intervals or bins. It displays the frequency or count of data points within specified ranges, with bars representing the frequency of data falling into each bin. The horizontal axis denotes the bins or intervals, while the vertical axis shows the frequency or count of data points. Histograms are particularly useful for understanding the distribution, variability, and central tendencies of a dataset.

Applications of Histogram Charts:

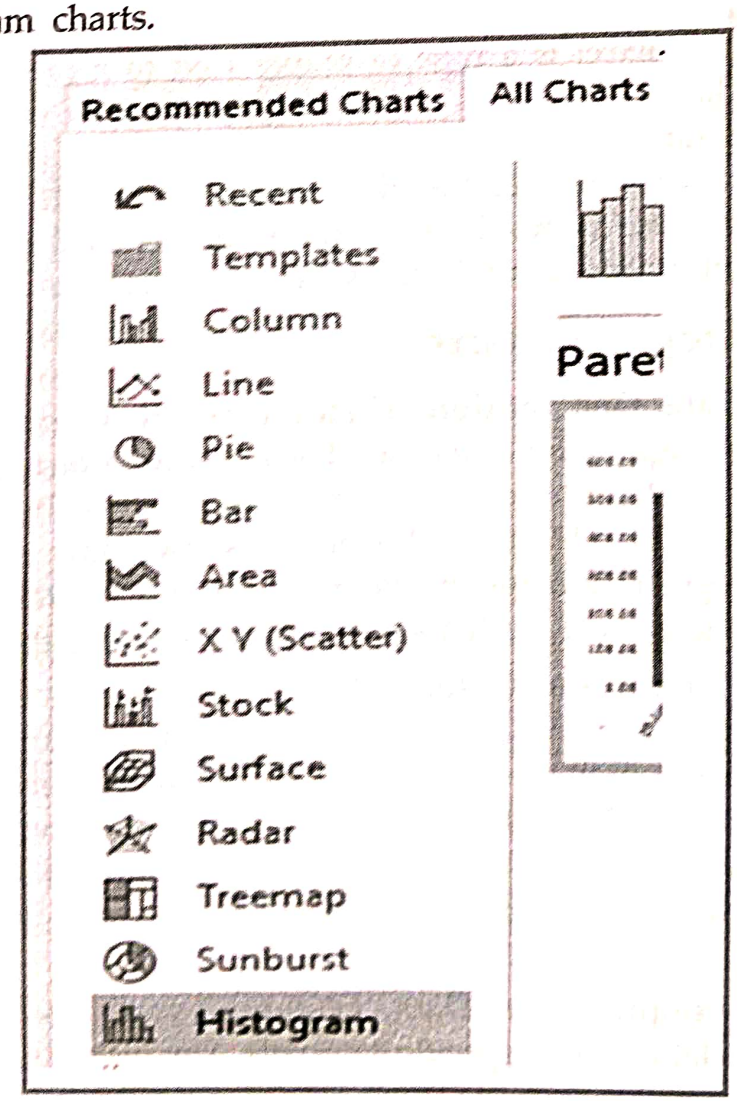
1. **Analyzing Data Distribution**: Histograms are valuable for analyzing the distribution of data within a dataset. For instance, a histogram could be used to visualize the distribution of exam scores for a class. By plotting the scores in intervals (e.g., 0-10, 11-20), the histogram shows how many students fall into each score range. This helps in understanding the overall distribution of scores, identifying where most students score, and assessing the spread of the data.
2. **Identifying Patterns and Outliers**: Histograms help in identifying patterns and detecting outliers within data. For example, if a company tracks the number of defects per production batch, a histogram can reveal whether there is a normal distribution of defects or if there are unusual spikes indicating potential issues. This visualization aids in recognizing patterns such as clusters of high or low values and highlights any anomalies or outliers that may need further investigation.
3. **Visualizing Frequency of Data**: Histograms are useful for visualizing the frequency of data within specific intervals. For instance, if a retail store tracks

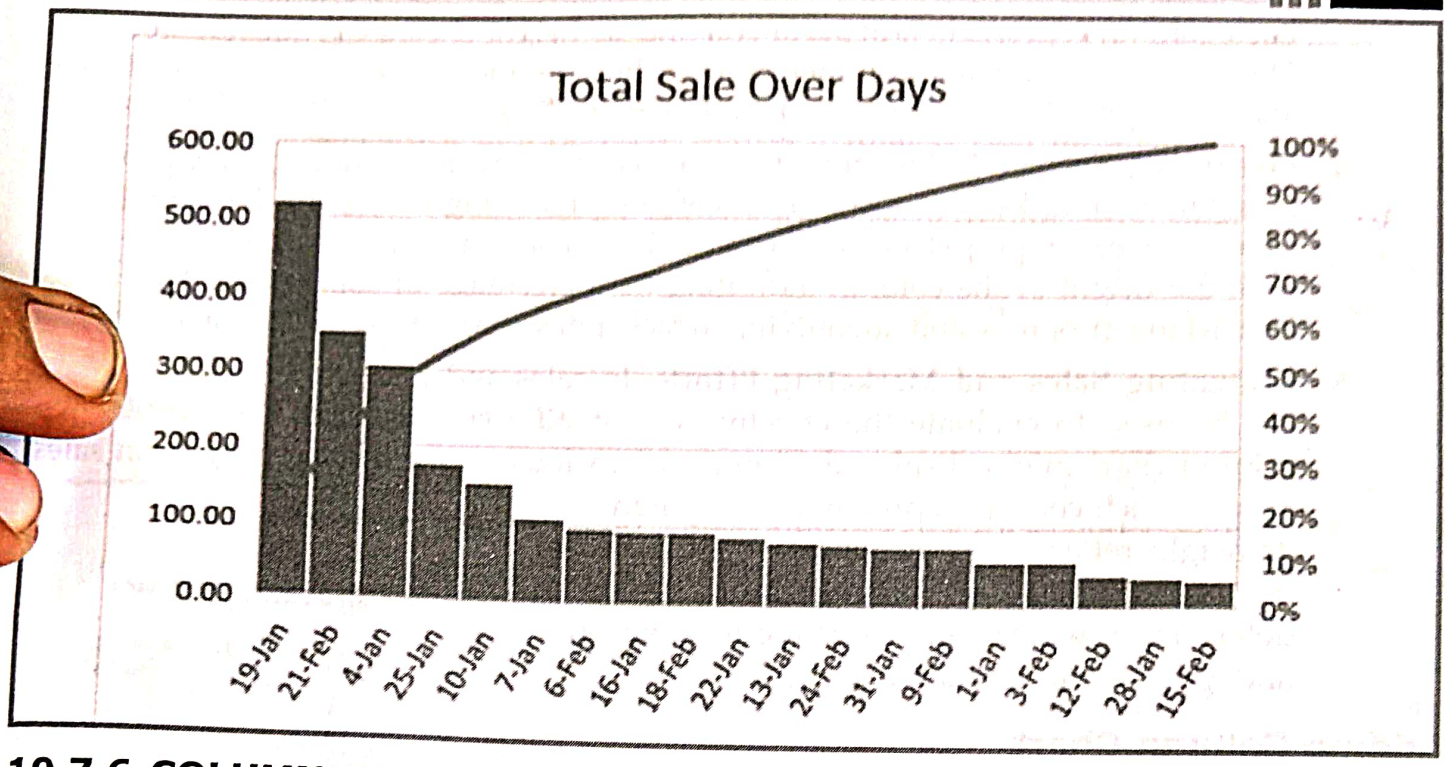
the number of sales made within different sales amounts (e.g., $0-$50, $51-$100), a histogram can show how frequently sales fall into these ranges. This application helps in understanding which sales ranges are most common and can assist in sales strategy and inventory management.

4. **Assessing Normality of Distribution:** Histograms can be used to assess whether data follows a normal distribution or another statistical distribution. For example, by plotting the height of individuals in a population, a histogram can reveal if the data approximates a bell-shaped curve, indicating a normal distribution. This assessment is important for statistical analysis and for determining appropriate statistical tests and methods.
5. **Determining Data Spread and Variability:** Histograms provide insights into the spread and variability of data. For example, a histogram of monthly temperatures over a year can show the range and frequency of temperature variations. By examining the shape and spread of the histogram, users can understand the degree of variability and identify periods with high or low variability in the data.

Adding Scatter Chart

Select the range and On the Insert tab, in the Recommended charts, select All charts and click on Histogram charts.





10.7.6 COLUMN CHART

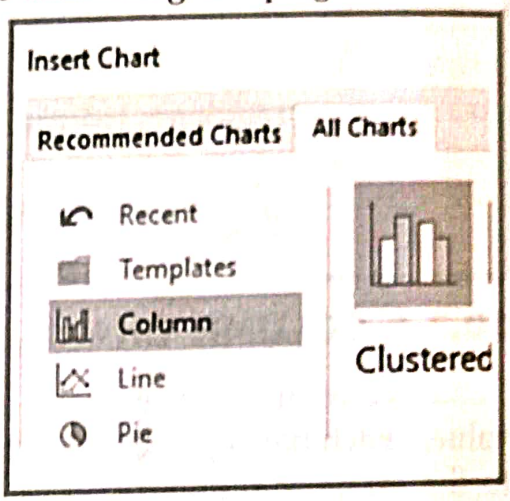
A column chart in Excel is a type of chart that uses vertical bars to represent data values. Each bar represents a category or a data series, and the height of the bar corresponds to the value it represents. Column charts are particularly useful for comparing different categories or data series across discrete intervals, such as months, regions, or product lines. The vertical arrangement of the bars makes it easy to visualize and compare data at a glance.

Applications of Column Charts:

1. **Comparing Categories**: Column charts are ideal for comparing different categories side by side. For example, a company might use a column chart to compare sales figures across different product lines. Each column represents a product line, with the height of the column indicating the sales volume. This allows for a clear comparison of sales performance across various products, helping to identify which products are performing well and which are under performing.
2. **Tracking Performance Over Time**: Column charts can be used to track performance over time by plotting data points at regular intervals. For instance, a column chart might show monthly revenue figures for a year. Each column represents a month, with the height of the column reflecting the revenue for that particular month. This application helps in visualizing trends, such as seasonal fluctuations or overall growth, and facilitates time-based performance analysis.
3. **Visualizing Budget Allocations**: Column charts are effective for visualizing budget allocations across different departments or projects. For example, a column chart could display the budget assigned to various departments within a company.

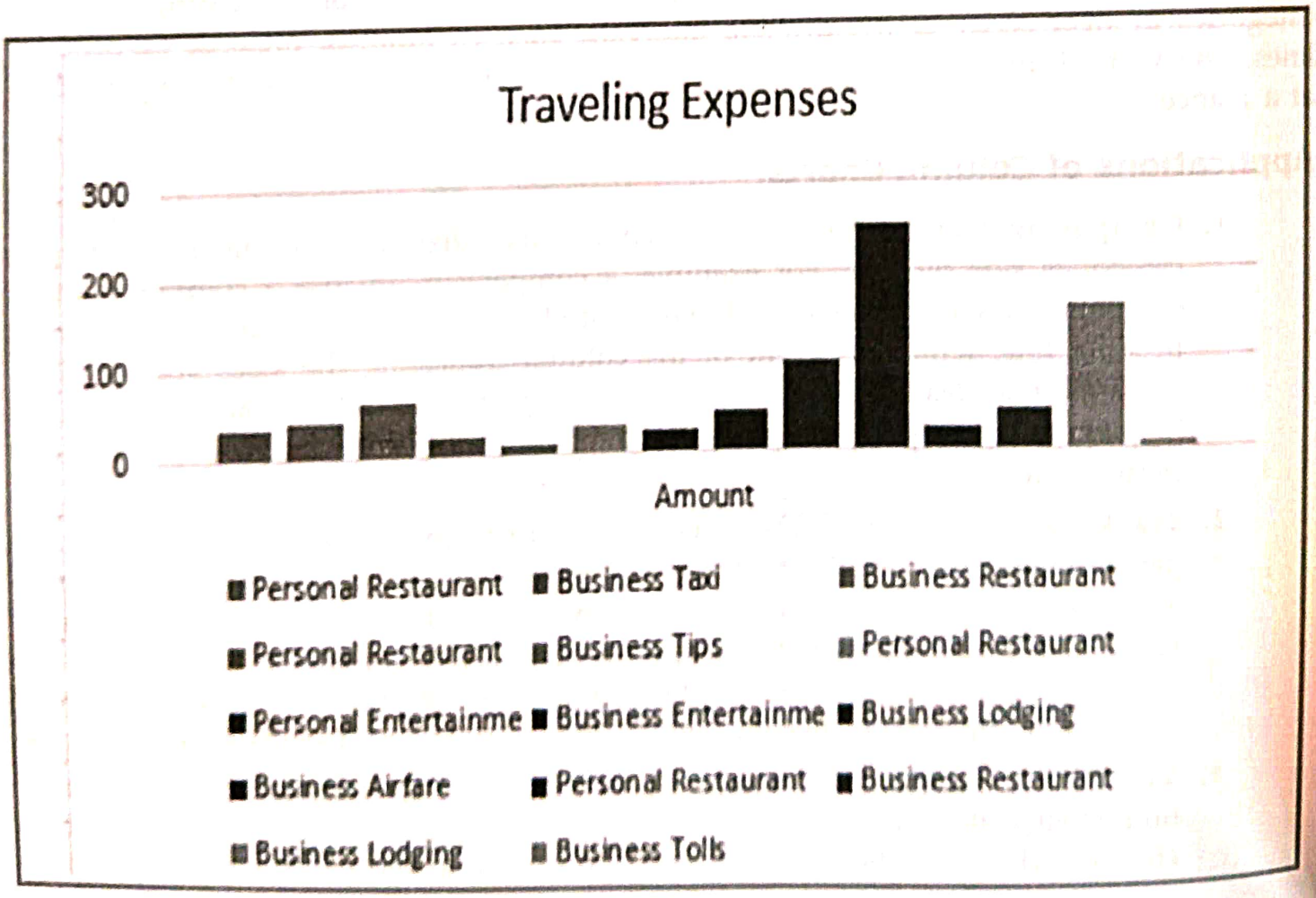
Each column represents a department, and its height shows the budget allocated. This visualization helps in understanding how financial resources are distributed and highlights areas with higher or lower funding.

4. **Presenting Project Milestones**: Column charts can be used to present project milestones or tasks. For instance, a column chart might display the completion status of various project phases or tasks. Each column represents a phase or task, with the height of the column indicating the percentage of completion. This helps in tracking progress and identifying which phases are on schedule or delayed.
5. **Evaluating Sales and Marketing Efforts**: In sales and marketing, column charts can be used to evaluate the effectiveness of different strategies. For instance, a column chart might display the impact of various marketing campaigns on sales figures. Each column represents a campaign, and its height reflects the sales increase attributed to that campaign. This visualization assists in determining which campaigns are most effective and guides future marketing efforts.



Adding Column Chart

Select the range and On the Insert tab, in the Recommended charts, select All charts and click on Column charts.



10.7.7 AREA CHART

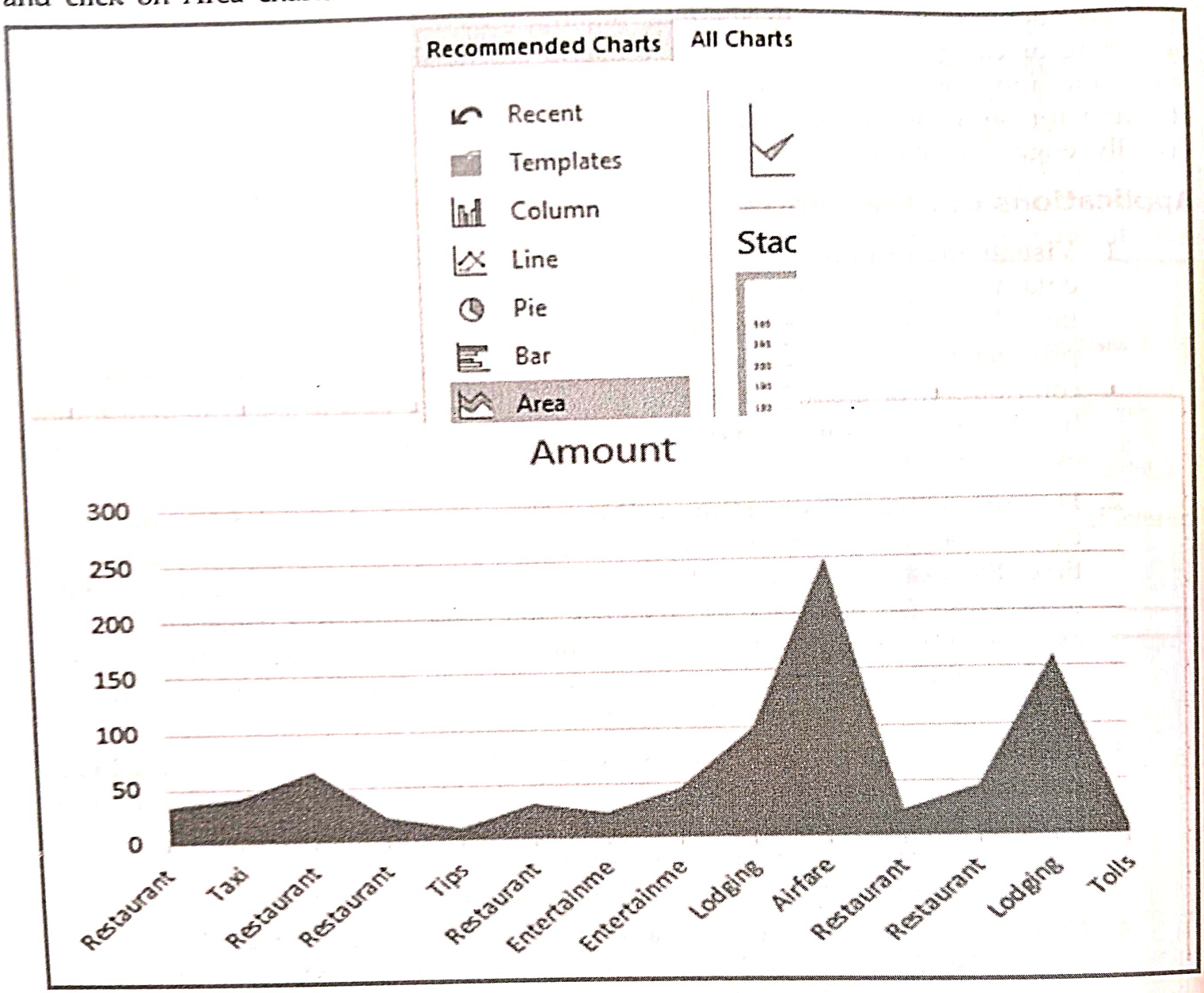
An area chart in Excel is a type of graph that displays data as a series of filled areas, typically beneath a line or lines, across a time period or other categories. The chart area is filled with color or shading, making it easy to see the magnitude of values over time or categories. Area charts are useful for illustrating how quantities change over time and how different data series contribute to a total. They are particularly effective for showing cumulative totals and the impact of multiple data series in a visually engaging manner.

Applications of Area Charts:

1. **Visualizing Cumulative Data**: Area charts are excellent for visualizing cumulative data, where you want to show how different categories contribute to a total over time. For example, an area chart can depict cumulative sales figures for various product lines over several months. Each area under the line represents the contribution of a product line to the overall sales. This helps in understanding how each product line's performance affects the total sales and how contributions evolve over time.
2. **Demonstrating Proportional Data Over Time**: Area charts effectively demonstrate how different categories contribute proportionally to a total over time. For example, an area chart could show the distribution of different types of expenses (such as salaries, materials, and overheads) as a percentage of total expenses throughout a year. By examining the areas under each line, users can see how the proportions of each expense category fluctuate over time and how they compare to the overall budget.
3. **Highlighting Trends and Patterns**: Area charts are useful for highlighting trends and patterns in data, especially when dealing with time-series data. For instance, an area chart might display the number of website visitors over several months. The filled areas make it easy to observe trends such as increasing or decreasing visitor numbers and to identify any seasonal patterns or anomalies in the data.
4. **Illustrating Market Share Dynamics**: Area charts can illustrate how market shares of different competitors evolve over time. For example, an area chart can show the market share of various companies within an industry, with each company represented by a different colored area. This helps in visualizing shifts in market dominance and understanding how each competitor's market share changes over time.
5. **Displaying Seasonal Variations**: Area charts are effective for displaying seasonal variations in data. For instance, a retailer might use an area chart to show sales figures throughout different seasons of the year. By filling the areas beneath the lines, the chart clearly illustrates seasonal peaks and troughs, helping to understand how sales fluctuate with seasons and plan inventory and marketing strategies accordingly.

Adding Area Chart

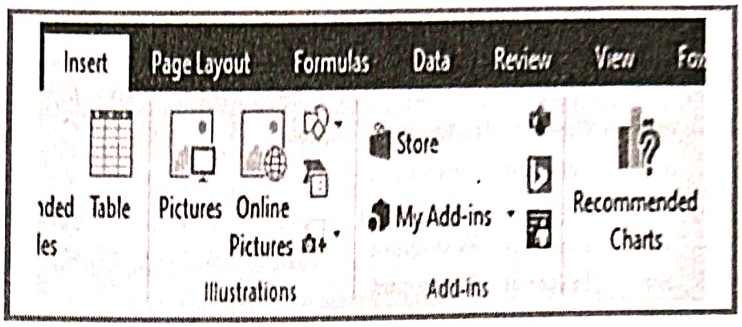
Select the range and On the Insert tab, in the Recommended charts, select All charts and click on Area charts.



10.8 CREATING EMBEDDED CHART

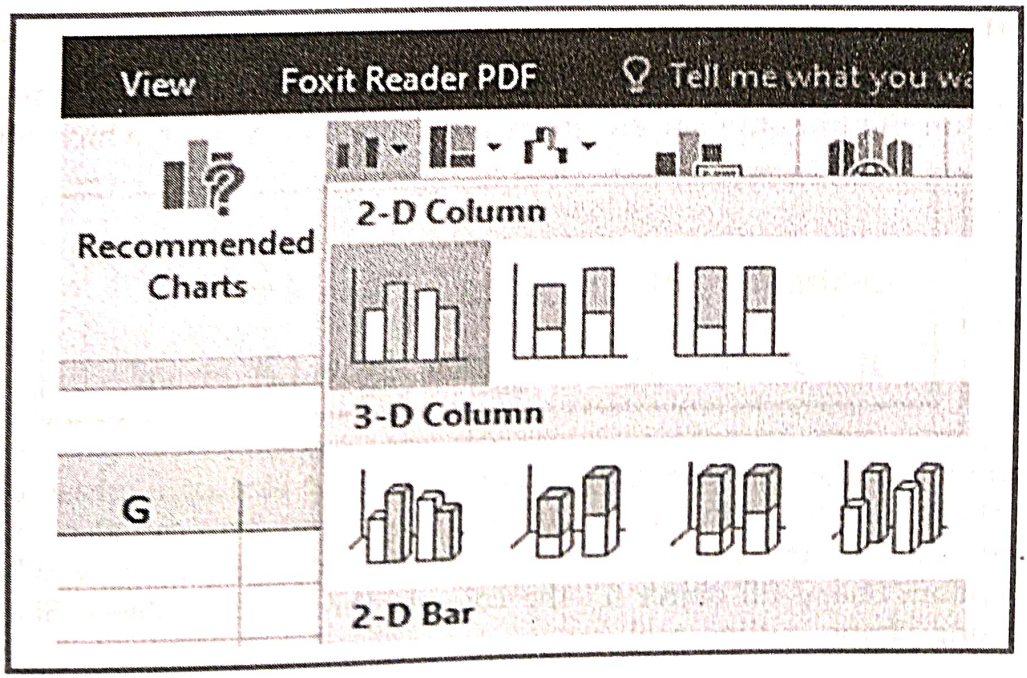
An embedded chart is a chart that appears right within the worksheet. Embed a chart on the worksheet when you want to be able to print the chart along with its supporting worksheet data. Keep in mind that all charts are dynamically linked to the worksheet that they represent. This means that if you modify any of the values that are plotted in the chart, Excel immediately redraws the chart to reflect the change, assuming that the worksheet still uses automatic recalculation. You can print any chart that you've embedded in a worksheet by itself without any worksheet data by selecting it before you open the Print dialog box.

1. Select range in worksheet and choose insert tab



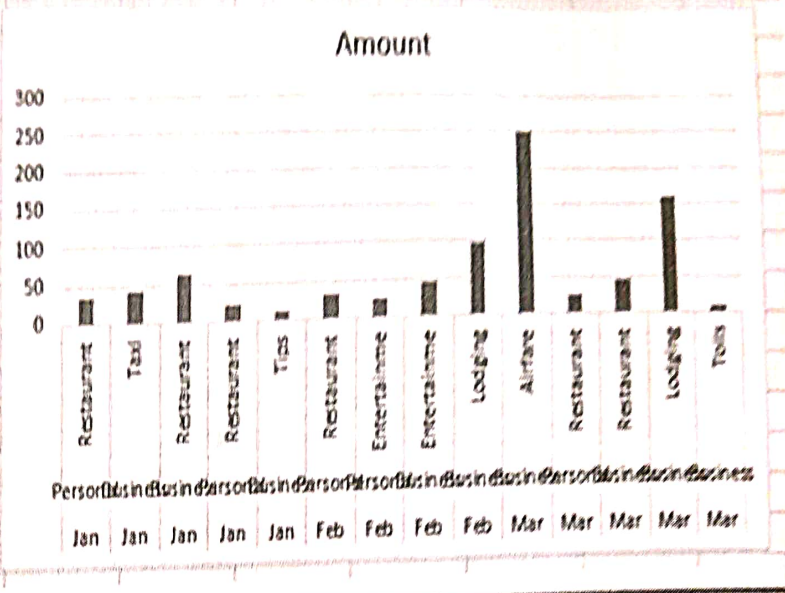
	A	B	C	D
1	Month	Types	Category	Amount
2	Jan	Personal	Restaurant	35
3	Jan	Business	Taxi	42
4	Jan	Business	Restaurant	64
5	Jan	Personal	Restaurant	22
6	Jan	Business	Tips	12
7	Feb	Personal	Restaurant	32
8	Feb	Personal	Entertainme	24
9	Feb	Business	Entertainme	45
10	Feb	Business	Lodging	98
11	Mar	Business	Airfare	250
12	Mar	Personal	Restaurant	24
13	Mar	Business	Restaurant	45
14	Mar	Business	Lodging	160
15	Mar	Business	Tolls	8

2. Click on chart type you want to create.



3. The chart will display on the same worksheet.

	A	B	C	D
1	Month	Types	Category	Amount
2	Jan	Personal	Restaurant	35
3	Jan	Business	Taxi	42
4	Jan	Business	Restaurant	64
5	Jan	Personal	Restaurant	22
6	Jan	Business	Tips	12
7	Feb	Personal	Restaurant	32
8	Feb	Personal	Entertainme	24
9	Feb	Business	Entertainme	45
10	Feb	Business	Lodging	98
11	Mar	Business	Airfare	250
12	Mar	Personal	Restaurant	24
13	Mar	Business	Restaurant	45
14	Mar	Business	Lodging	160
15	Mar	Business	Tolls	8



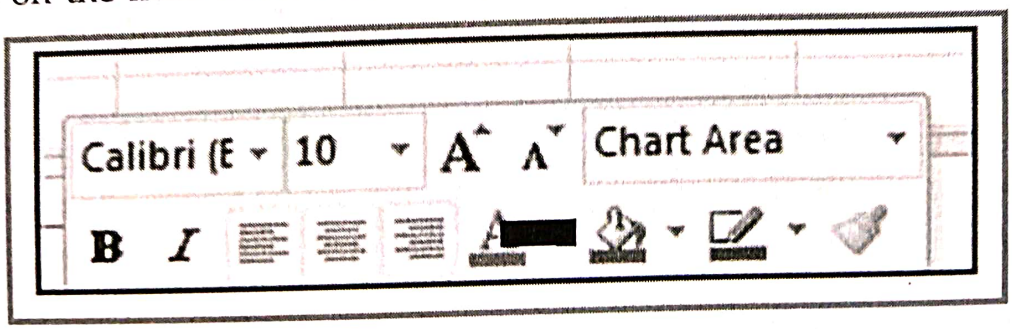
10.9 FORMATTING CHART

You can format each individual element that makes up a chart. This helps to enhance the appearance of each element so that the chart displays your data in the most effective manner possible. Depending on the item you select, the options vary, and only the most common are described here. For chart titles, you can format the font, font style, font size, alignment, border, and pattern. Excel allows you to work with individual chart elements and apply additional formatting.

Followings are the different ways to format the chart.

Formatting Chart Using Mini Tool-Bar

The mini-toolbar provides a selection of popular formatting options. In fact, the options shown on the mini-toolbar are always the same, no matter which object you click.

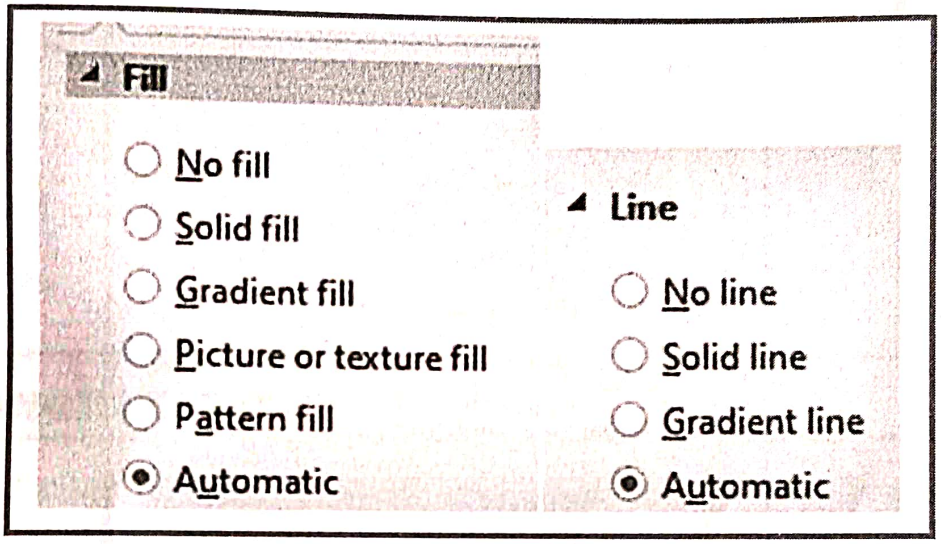


Using mini Tool-Bar you can change Font and Font size or increase or decrease the font size. Font can be align using three basic cell alignment. You can also change the color. You can select the different chart element using the drop down arrow which shows the 'Char area' option. Shape fill object fill the color in the selected object. Shape border can be colored using shape outline tool.

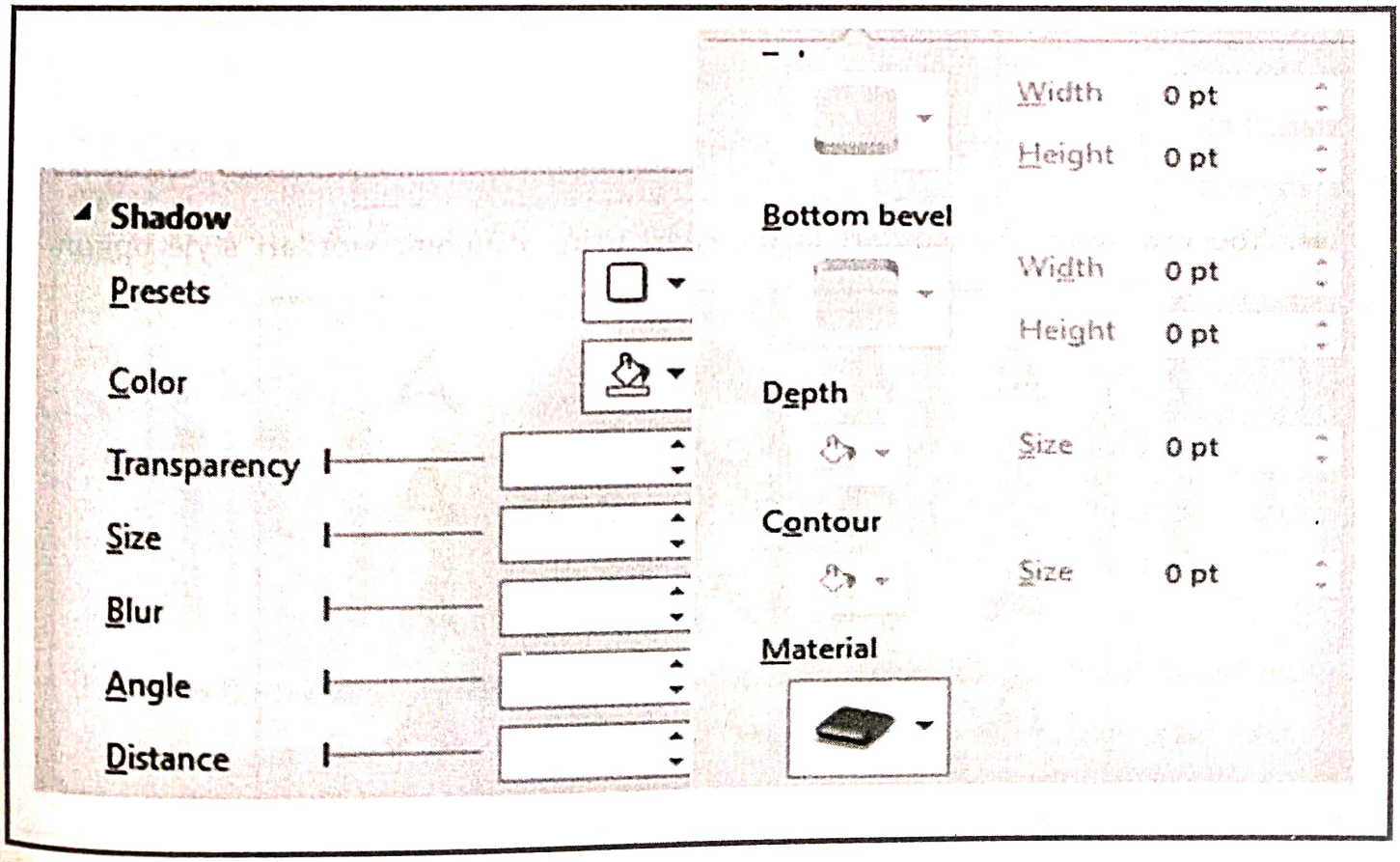
Formatting Using Format Tab

The contents of the format dialog box will vary with the object you've selected. The formatting options for data series, chart, and plot areas are basically of the change-their color variety.

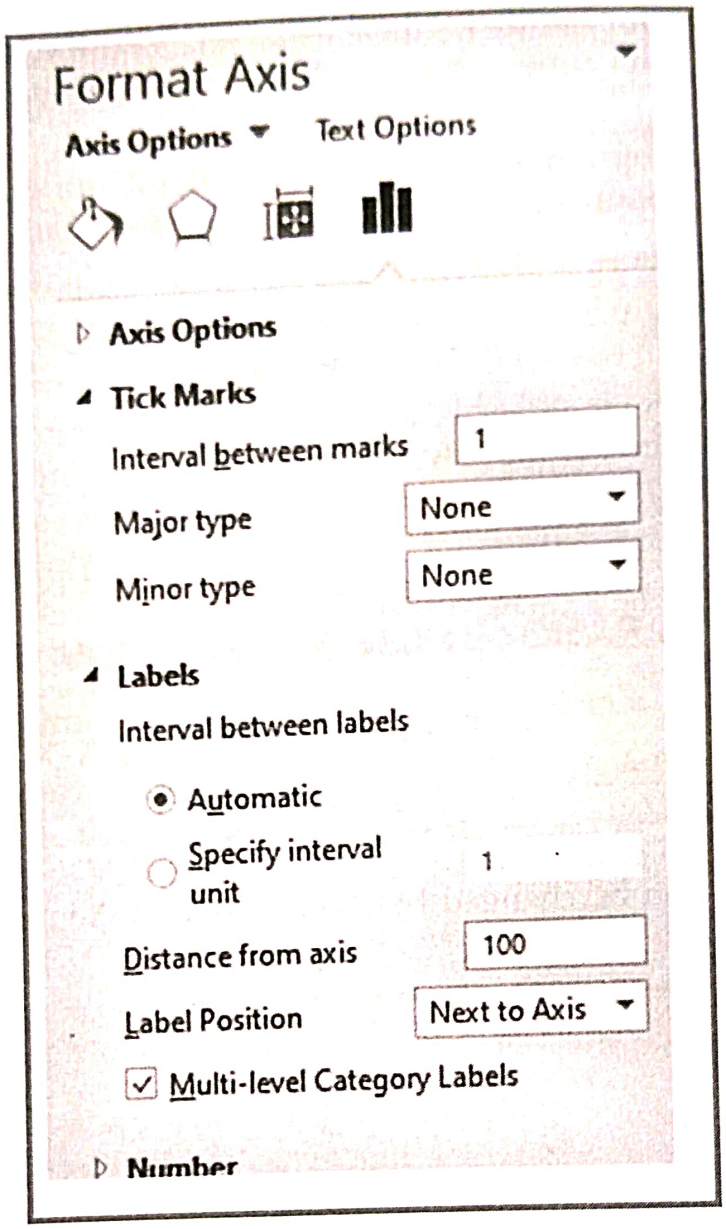
- Format has fill option which has different fill option for the chart you have selected. Line option draw different lines in the chart.



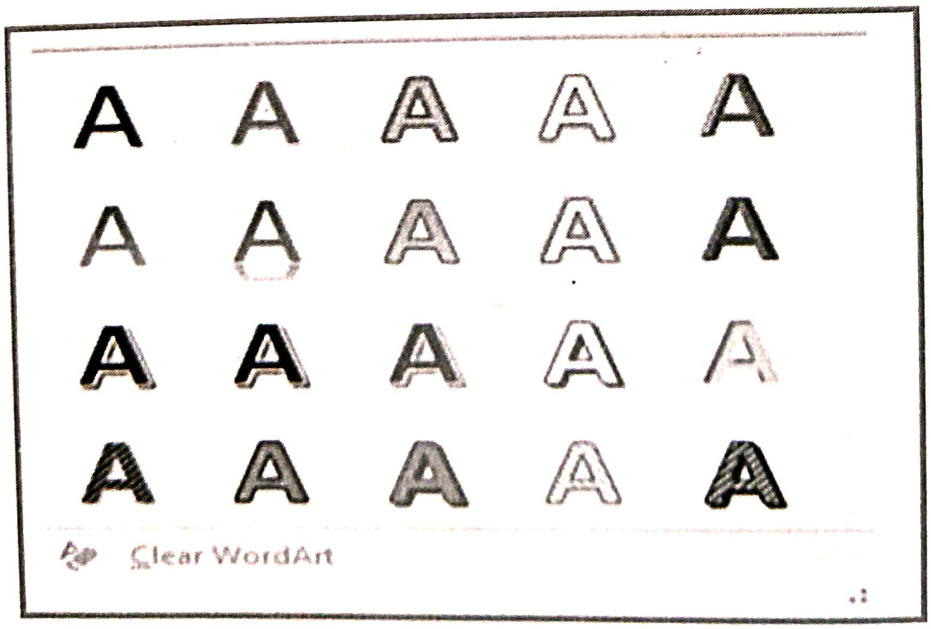
- Chart shadow can be changed by shadow option. You can apply the 3D effect to the 3D format.



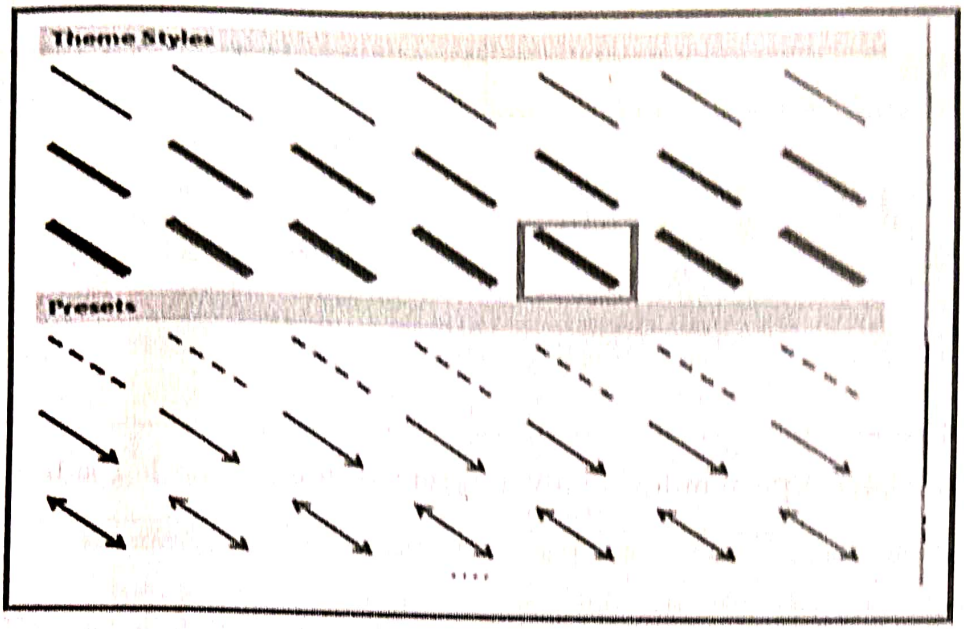
- Chart axis can be formatted using Format Axis



- You can apply the wordart to the chart using different wordart style option.



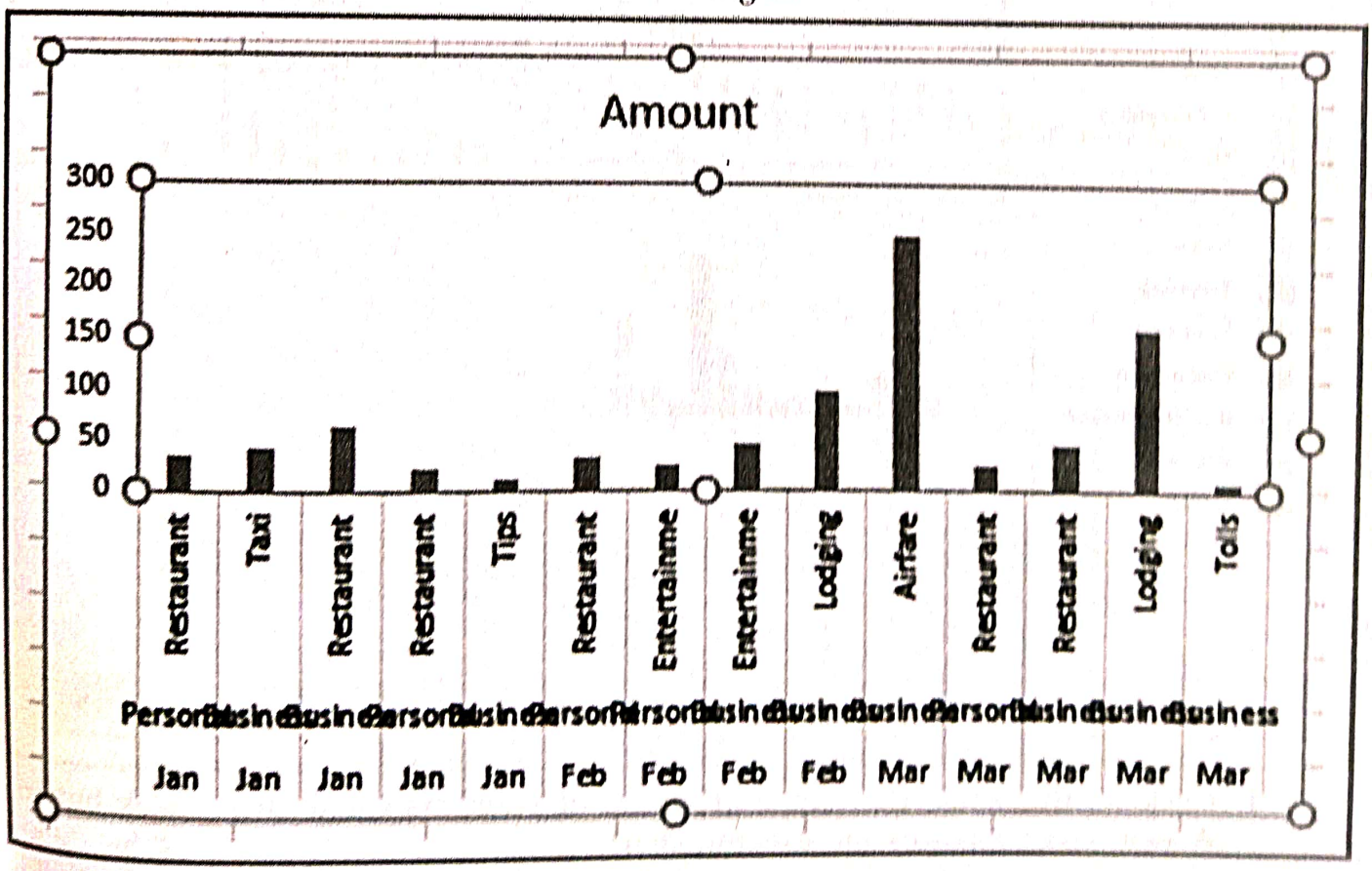
- Shape style option is used to apply on different shapes.



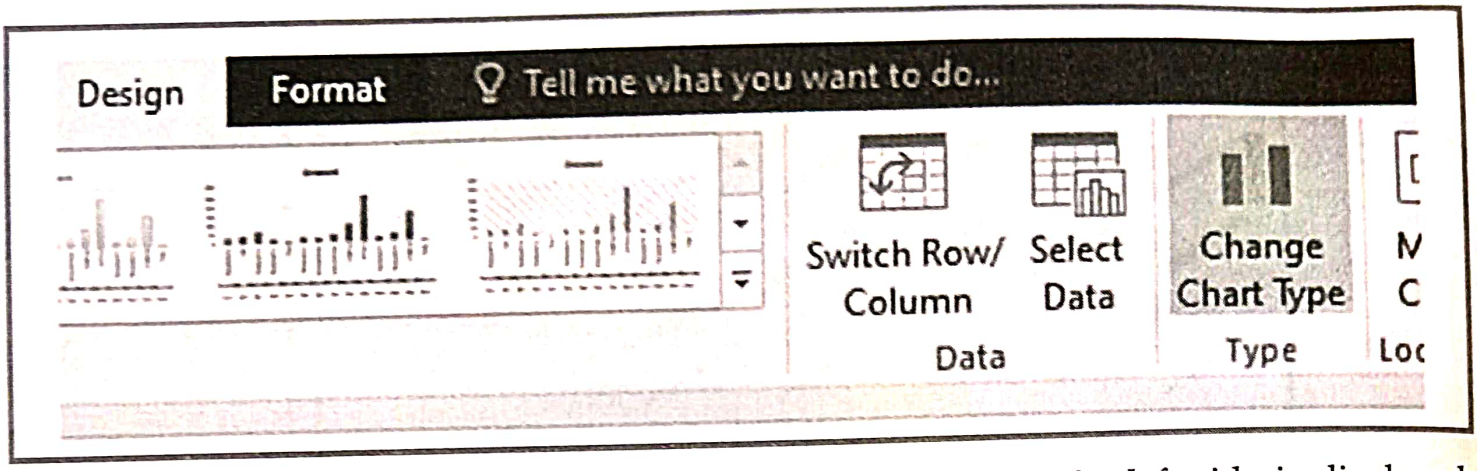
10.10 CHANGING CHART TYPE

In the visualization of data, different kinds of charts are used. You should move to some other chart that better explains your data if the chart format you used to display your data is not suitable or necessary.To complete this task, follow these steps:

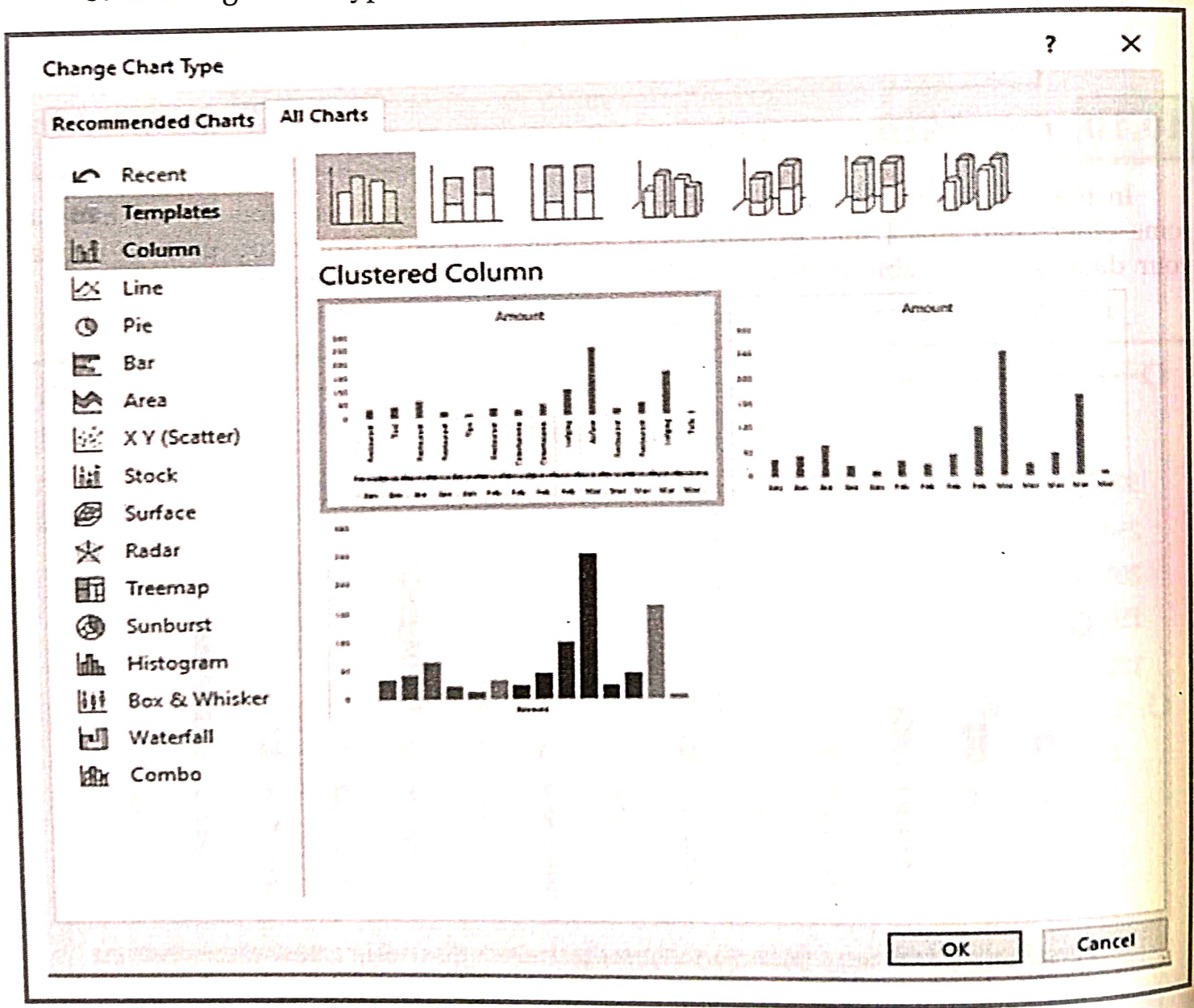
1. Select the chart which has to be changed.



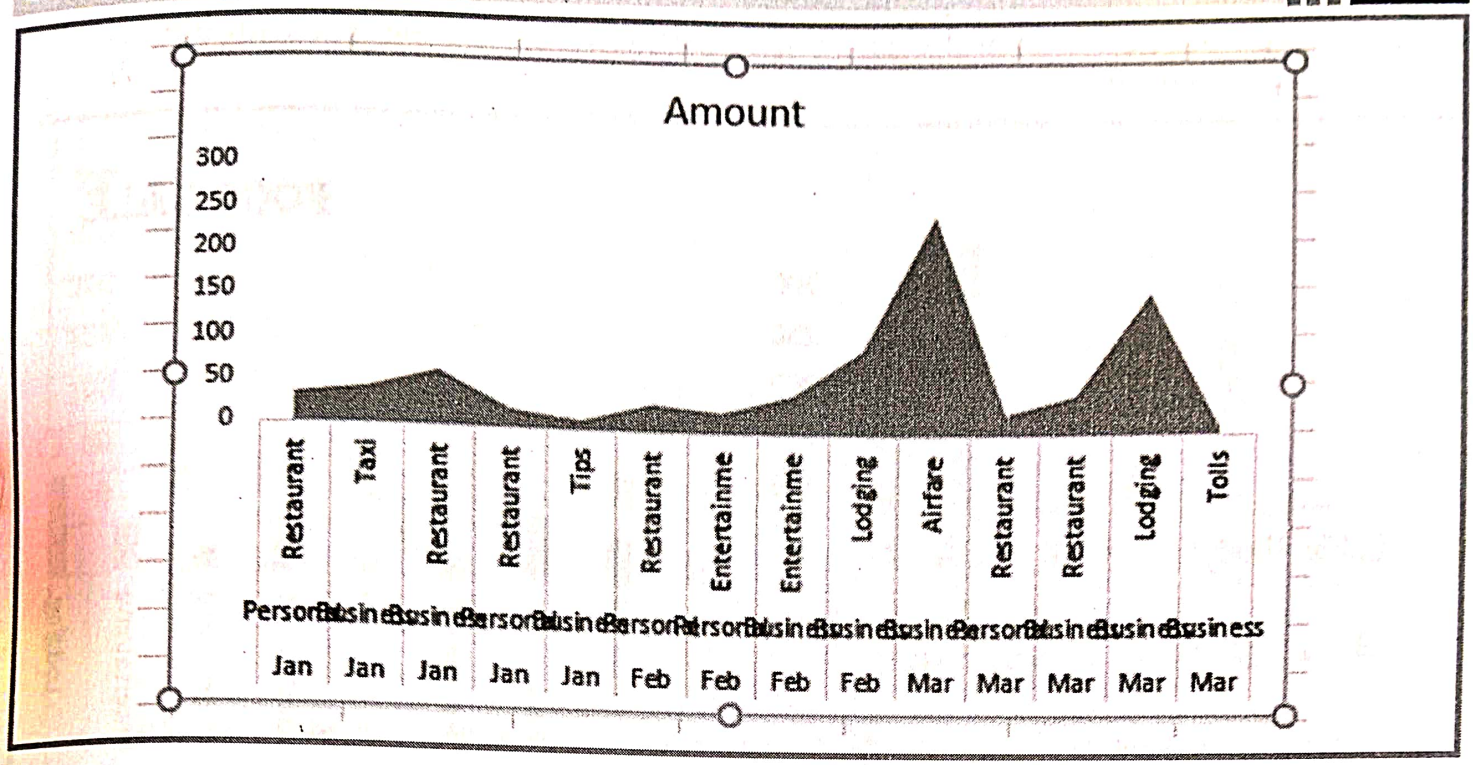
2. Click on design tab and the click on change chart type option.



3. A change chart type window with different option at the left side is displayed.



4. Click on the 'Area' chart type or choice of your chart and click on ok button. A new chart replaces the existing chart.

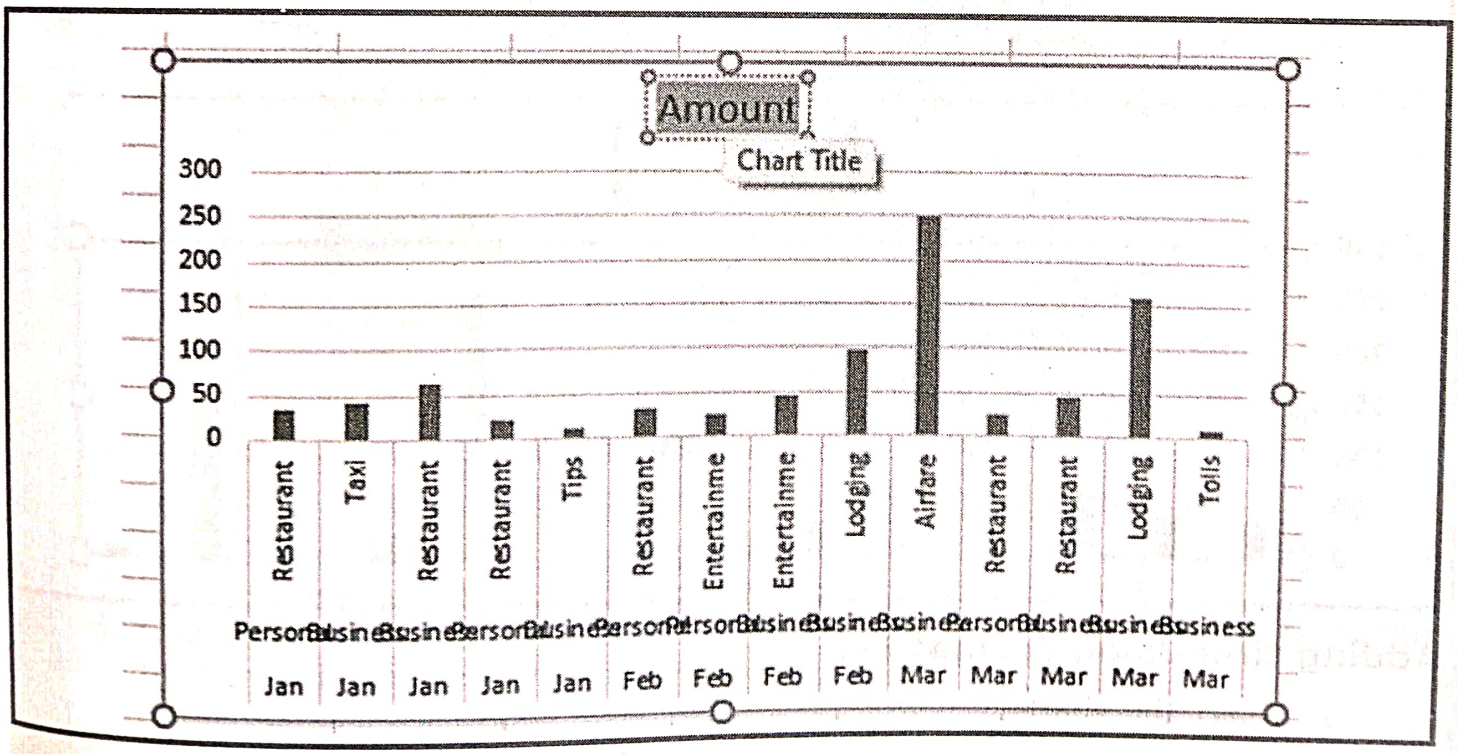


10.11 ADDING TITLE, LEGEND AND GRIDLINES

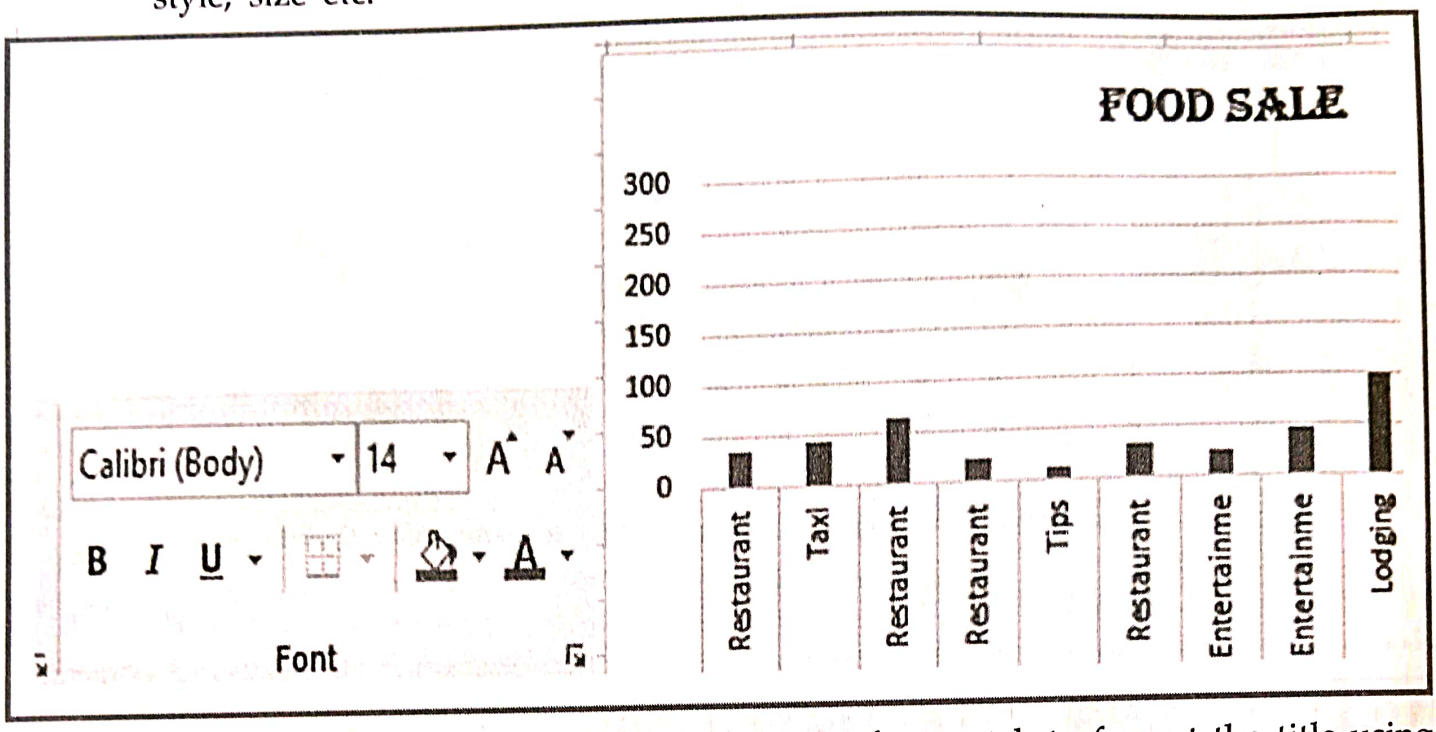
When Excel first adds any title to a new chart, the program gives it a generic name, such as Chart Title. or AXIS TITLE. To replace such generic titles with the actual chart titles following steps are used.

Adding title using Editing text

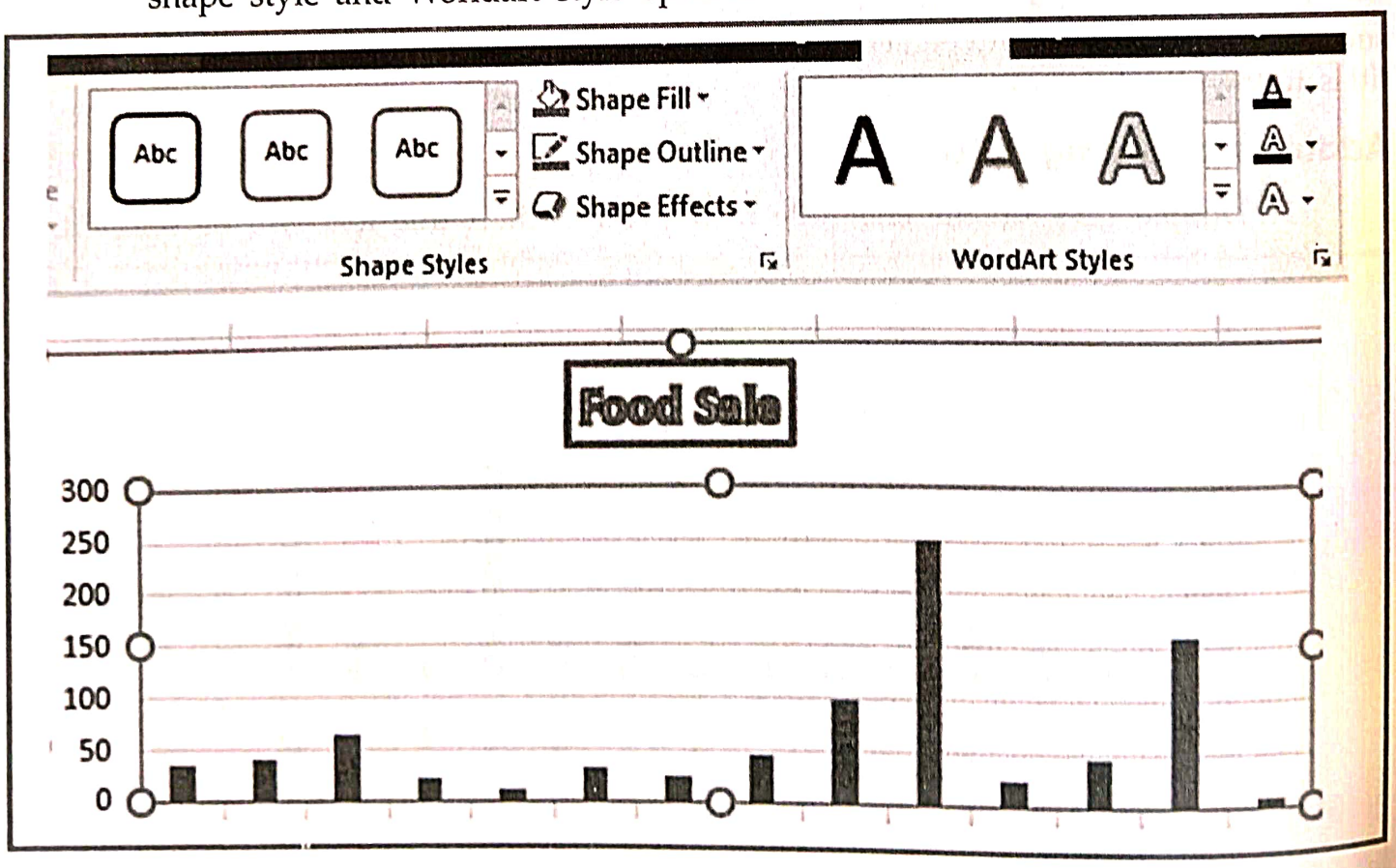
1. Double click on the char title in the chart. Chart title text will be selected.



2. Type the new title and click on the font tab to format the title with different font style, size etc.

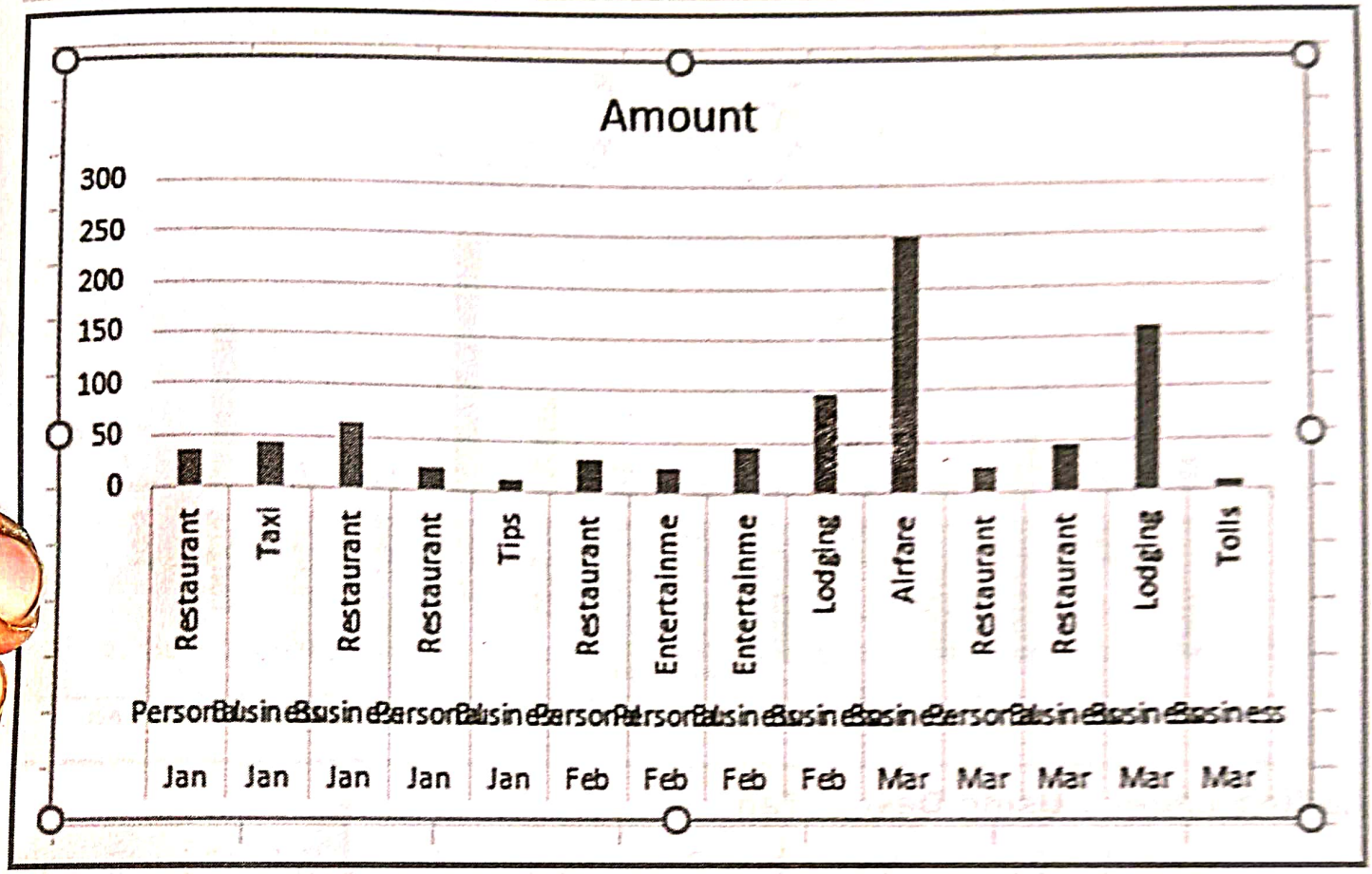


3. To apply special effect to chart title, click on format tab to format the title using shape style and Worldart style options.

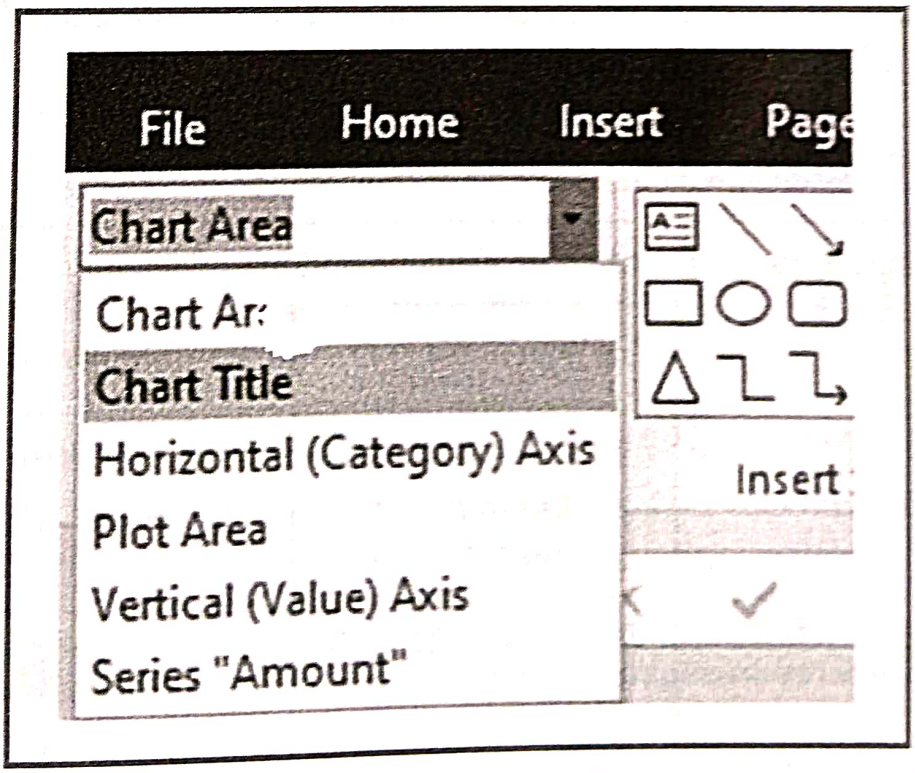


Adding title Using Format Tab

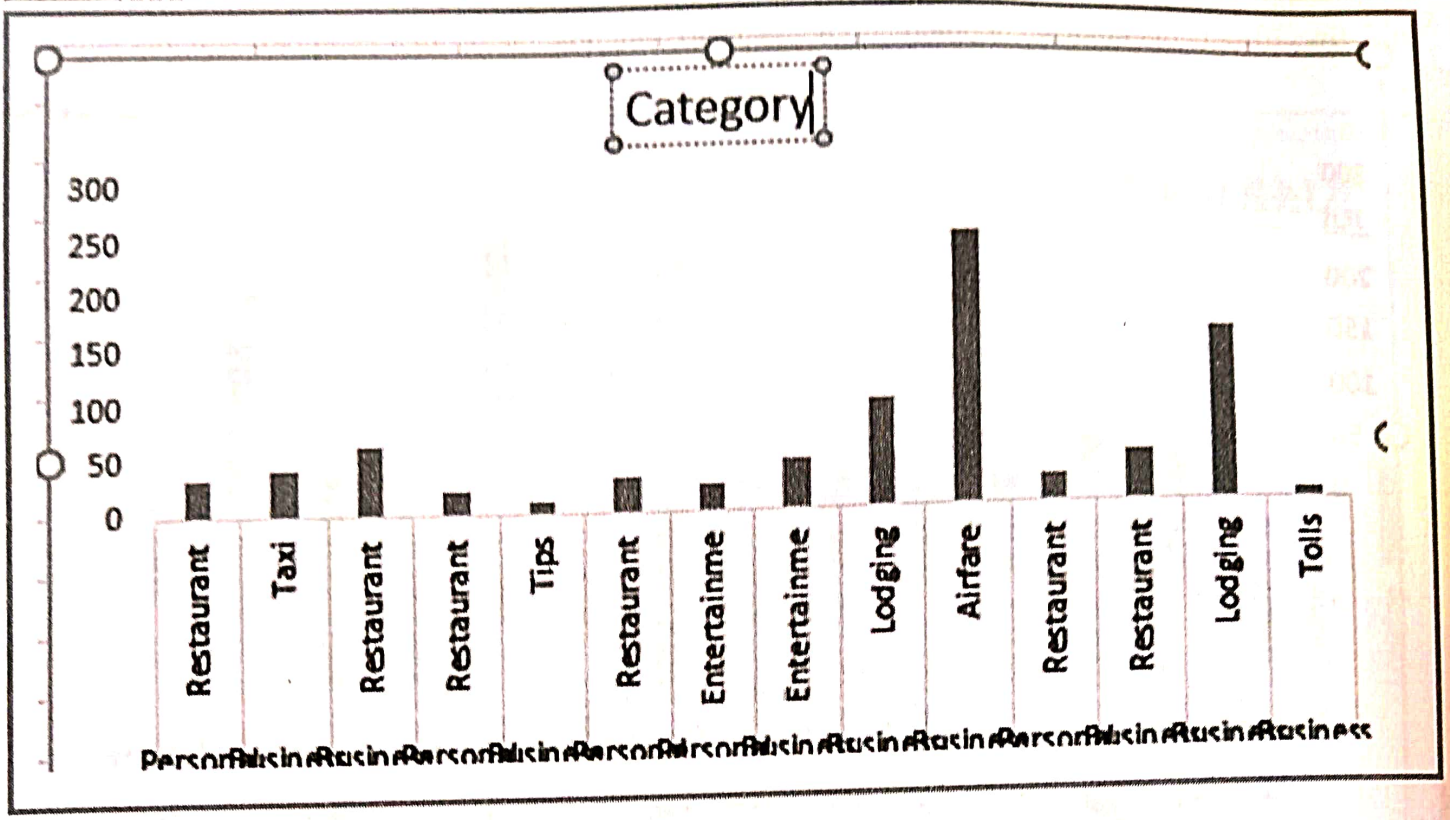
1. Select the chart.



2. Click on the format tab and select 'chart title' from the drop down at the left most in the format tab.

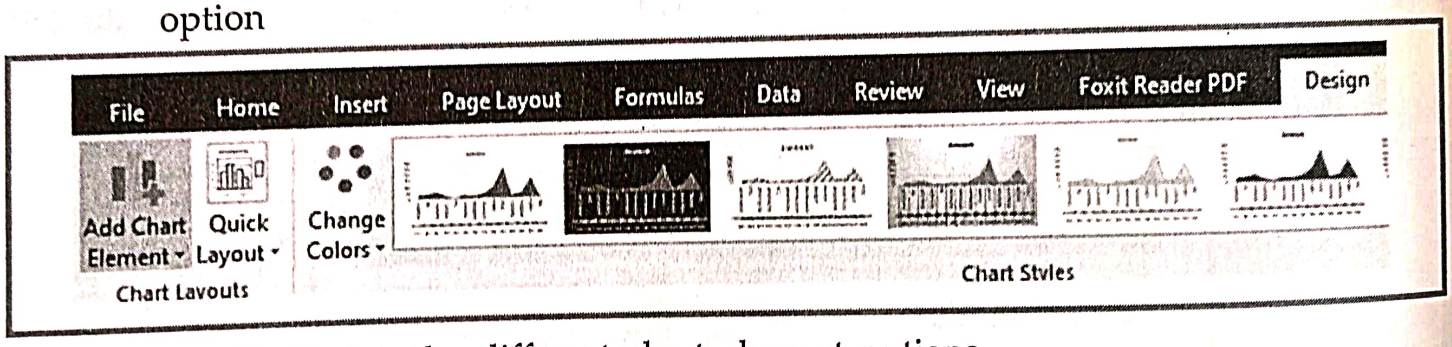


3. Chart title will be selected in the chart. Now you can enter the new chart title by typing in the selected text box.

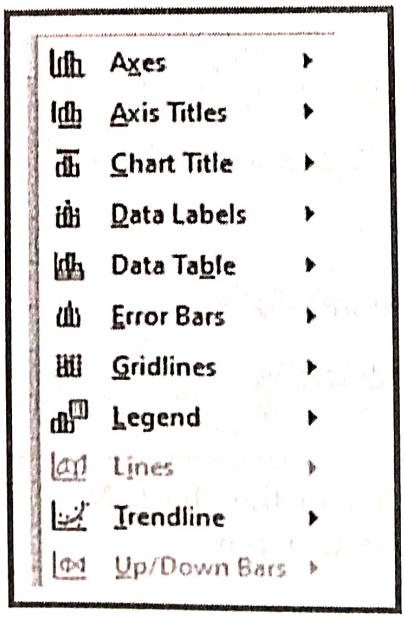


Adding Legend Using Design Tab

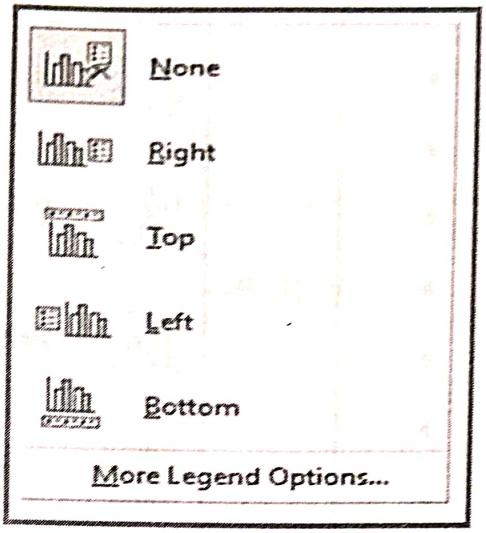
1. Select the chart and the click on Design tab then click on 'Add Chart Element' option



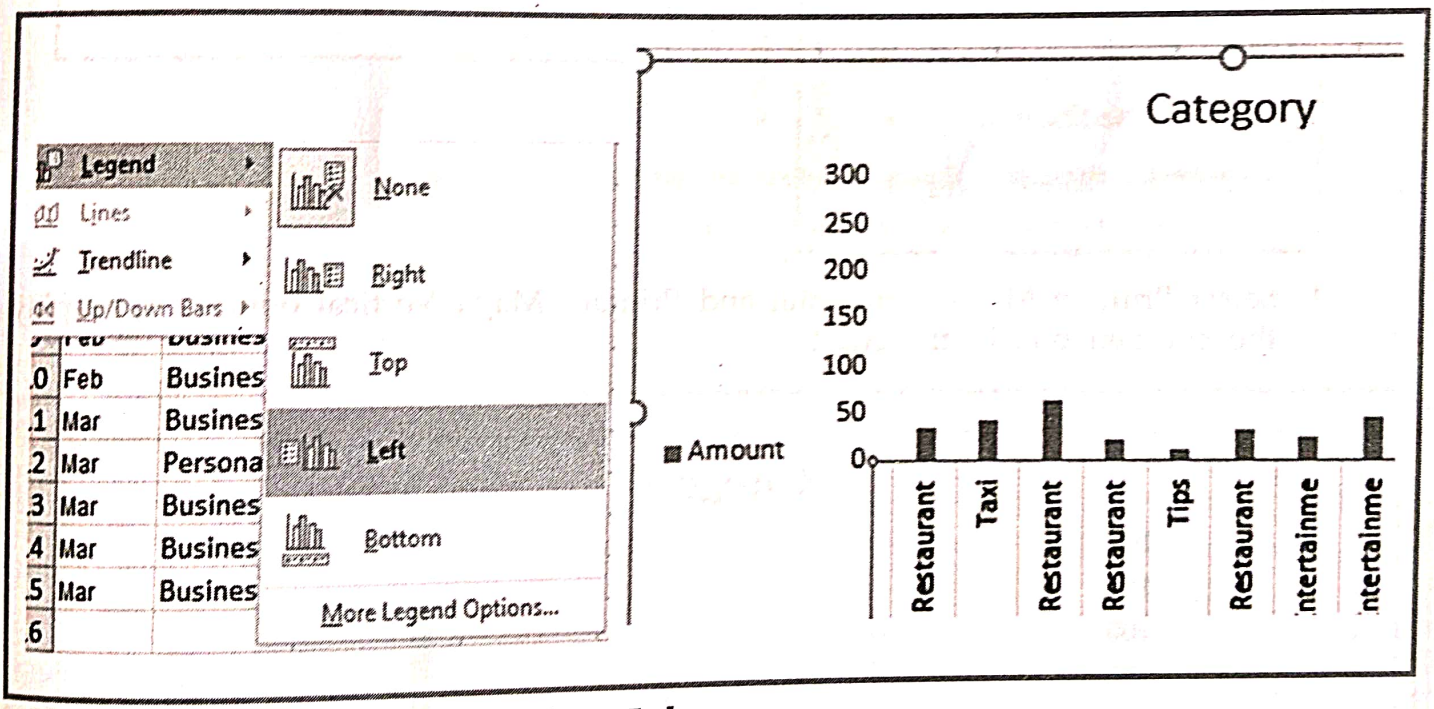
2. It will display the different chart element options.



3. Click on the Legend option. A different legend options are displayed to place the legend in the Top, left, right or bottom.

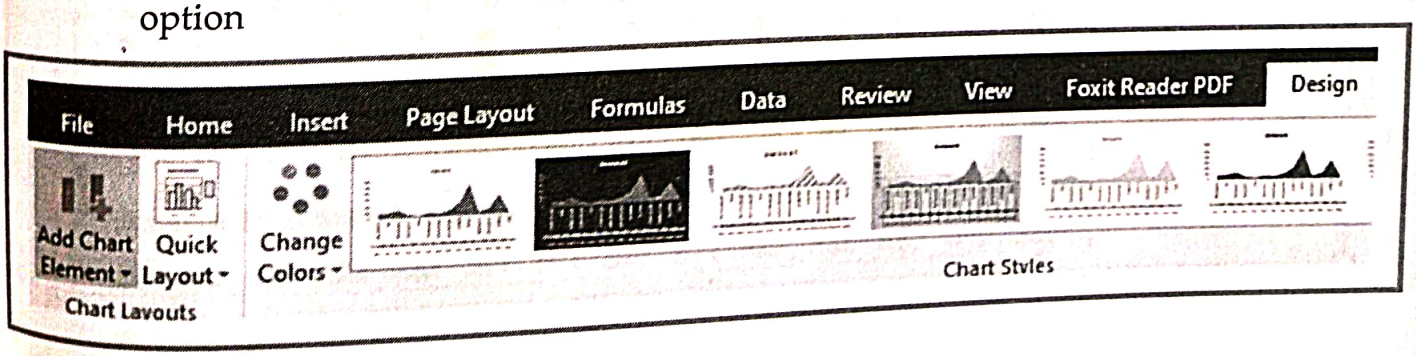


4. Select left options to place the legend at the left in the chart as shown below.

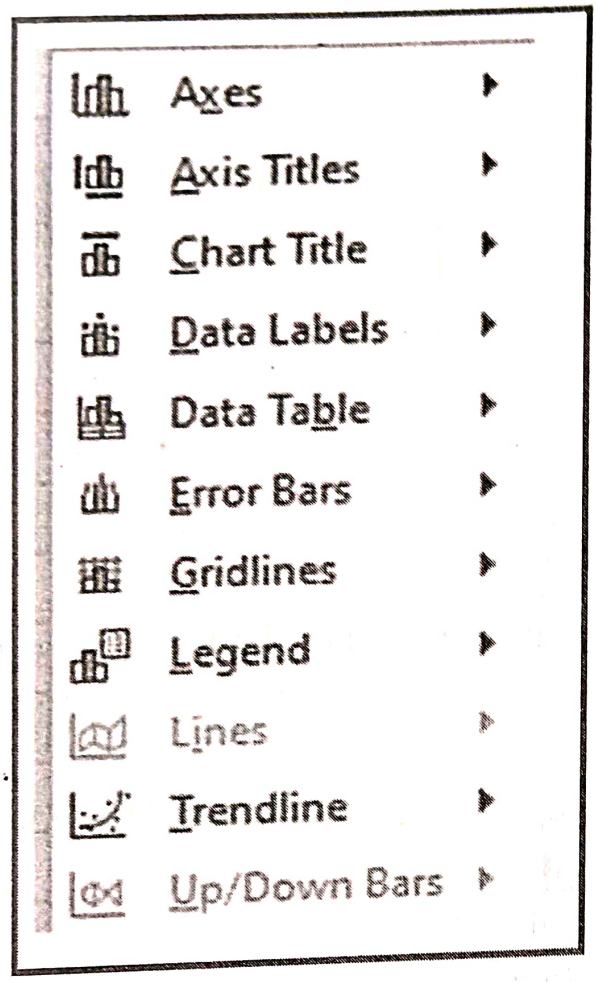


Adding Gridlines Using Design Tab

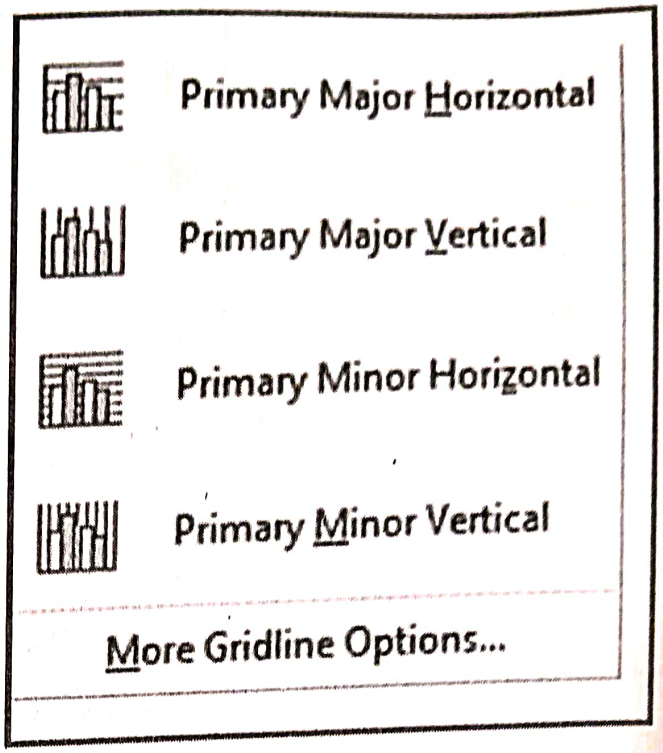
1. Select the chart and the click on Design tab then click on 'Add Chart Element' option



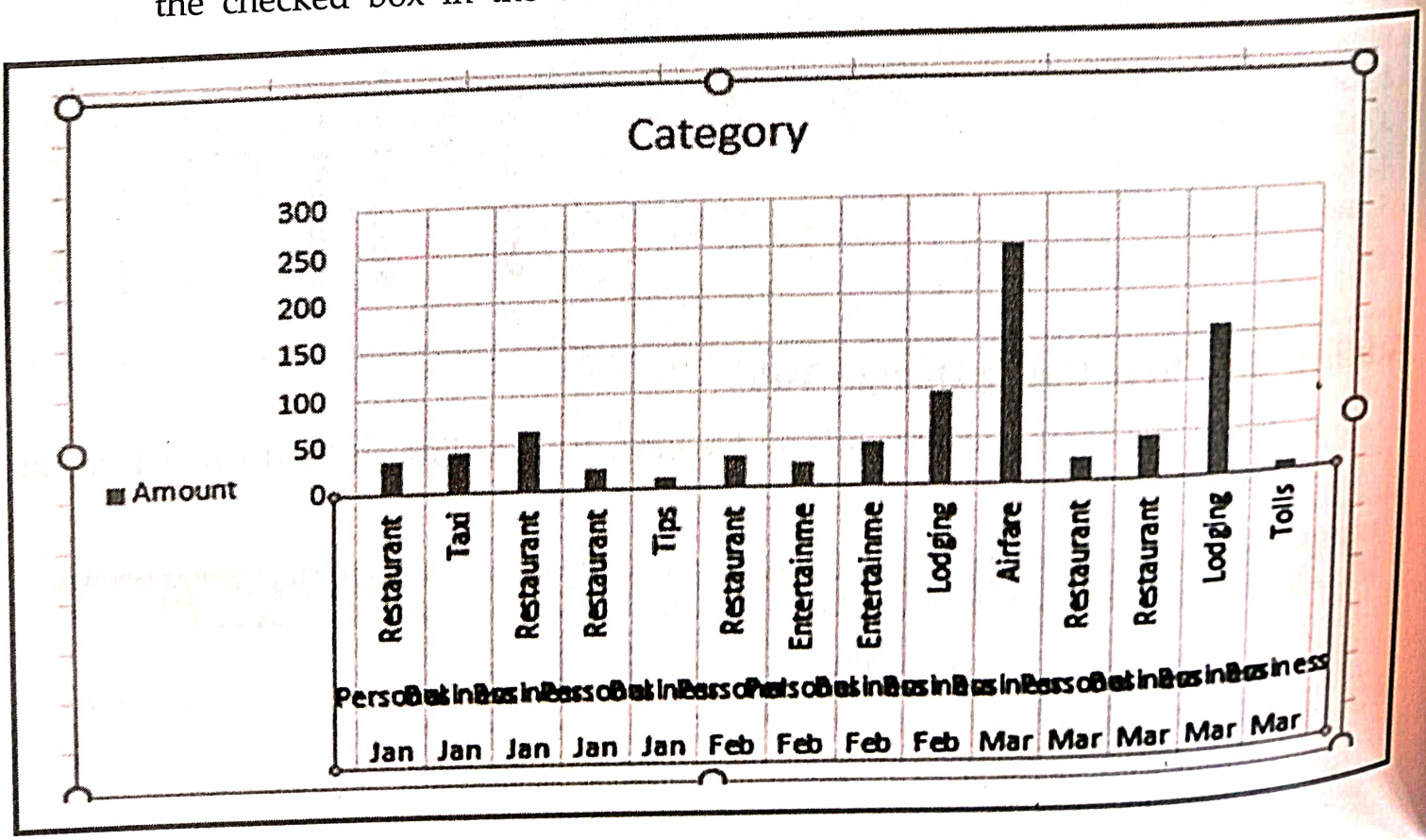
2. It will display the different chart element options.



3. Click on the Gridlines option then different gridlines options are appears.



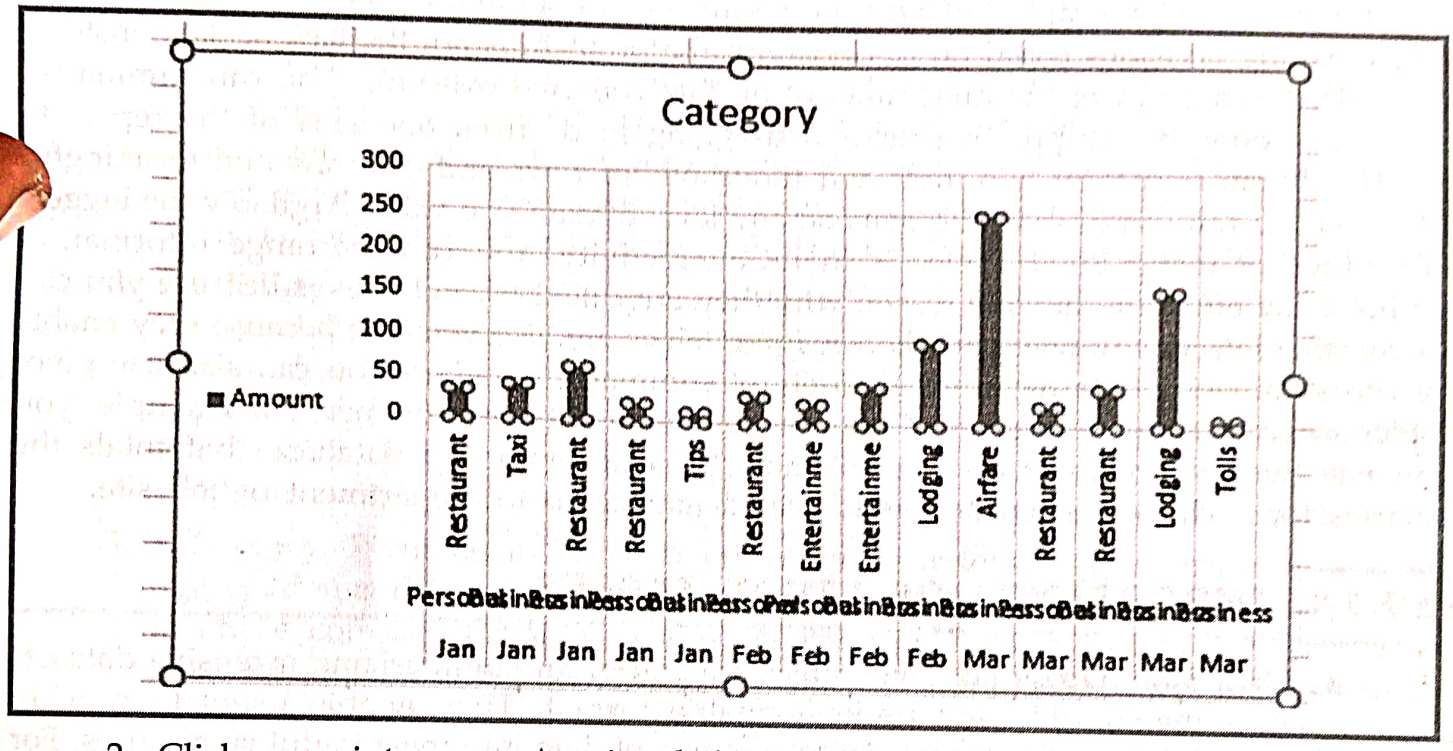
4. Select Primary Major Horizontal and Primary Major Vertical options to display the checked box in the chart.



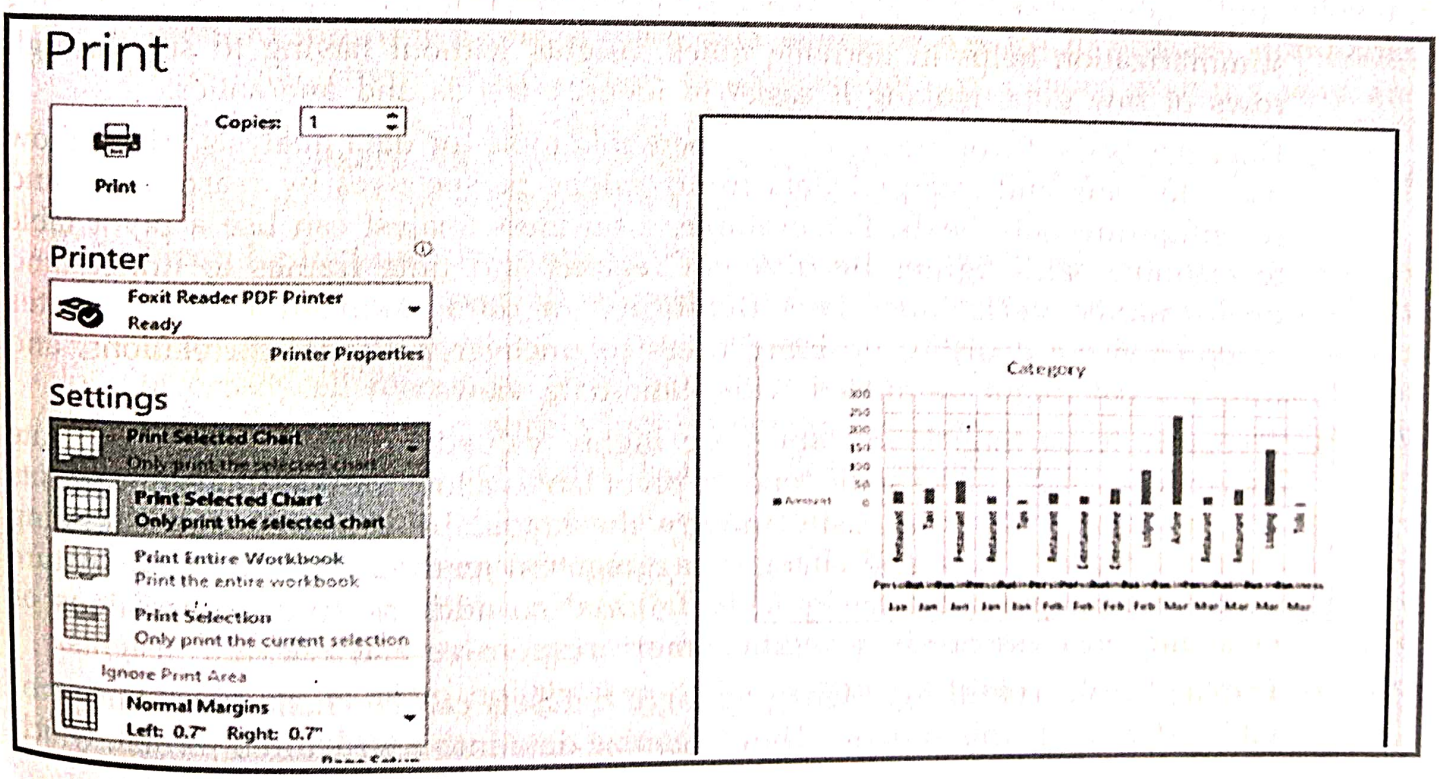
10.12 PRINTING CHART

When you have an embedded chart in your worksheet, it prints as positioned, along with the data, when you select *Print* from the *File* tab and click the *Print* button. To print the chart on its own

1. Select the chart and then click on File menu



2. Click on print menu. A print dialog box will be displayed. Under the Setting select Print Selected Chart



3. If chart does not display properly then use page break preview and reposition the chart in a single page
4. Click on print button to print the chart as you print normal worksheet.

10.13 INTRODUCTION TO PIVOT TABLE

Pivot table is a simple but very powerful tool to summarize and analyze data and helps making decisions based on it. Generally, the problem is not the lack of data; instead, it's making sense out of the huge amount of data collected overtime. You can summarize data in seconds by simply clicking and dragging fields from one area of the report to another. It can transform hundreds and thousands of rows and columns into meaningful information. You can display subtotals on any level of data you want. Arguably the biggest advantage of using Pivot Tables is in its interactivity. You can rearrange information without any effort. In fact, the reason why they are called Pivot Tables is because you can 'pivot' the data in any way you choose. Pivot tables are also versatile because they enable you to summarize data by using a variety of summary functions. You can also use pivot tables to cross-tabulate one set of data in your data list with another. For example, you can use this feature to create a pivot table from an employee database that totals the salaries for each job category cross-tabulated (arranged) by department or job site.

10.14 IMPORTANCE OF PIVOT TABLES

1. **Data Summarization:** Pivot tables are crucial for summarizing extensive datasets in a manageable and understandable format. They enable users to quickly consolidate and aggregate large amounts of data into meaningful summaries. For instance, if you have a spreadsheet with thousands of sales transactions, a pivot table can instantly calculate total sales by product, region, or time period. This summarization helps in deriving quick insights without having to sift through rows of raw data, making it easier to identify trends and anomalies.
2. **Data Analysis:** Pivot tables are indispensable tools for data analysis. They allow users to view and interpret data from various perspectives by rearranging and reconfiguring data fields. For example, a business analyst can use a pivot table to examine sales figures by different regions and time frames to understand performance variations. This flexibility in data manipulation facilitates comprehensive analysis, enabling users to uncover patterns, correlations, and insights that might be missed with static data views.
3. **Dynamic Reporting:** Pivot tables are highly valued for their ability to create dynamic reports. Unlike static reports, pivot tables allow for real-time adjustments and updates. Users can easily modify the layout by dragging and dropping fields, applying filters, and changing aggregation methods. This dynamic nature means that reports can be quickly tailored to address specific questions or changing business needs, providing more responsive and relevant insights.
4. **Efficient Data Handling**: Managing large datasets can be challenging, but pivot tables streamline this process. They organize data into a structured format, which

simplifies complex datasets into more digestible summaries. For instance, a pivot table can aggregate data from multiple sheets or sources, presenting a unified view that is easier to analyze and interpret. This efficiency reduces the manual effort required for data manipulation and enhances overall productivity.

10.15 ADVANTAGES OF PIVOT TABLES

1. **Ease of Use:** Pivot tables are designed to be user-friendly, even for those with limited technical skills. They rely on intuitive drag-and-drop functionality rather than complex formulas. Users can create a pivot table by simply selecting data ranges and arranging fields into rows, columns, and values. This ease of use makes pivot tables accessible to a wide range of users, from beginners to advanced analysts, facilitating broader data analysis capabilities across different skill levels.
2. **Flexibility:** The flexibility of pivot tables is one of their greatest strengths. Users can quickly pivot data by rearranging rows and columns to view information from different angles. They can also apply various aggregation functions such as sum, average, count, and percentage to analyze data in diverse ways. This adaptability allows users to customize their analysis according to their specific needs and explore different dimensions of the data with minimal effort.
3. **Efficiency:** Pivot tables enhance efficiency by automating the process of data aggregation and summarization. Instead of manually calculating totals or creating multiple complex formulas, users can leverage pivot tables to perform these tasks instantly. This efficiency not only saves time but also reduces the likelihood of errors that can occur with manual calculations. The automation of repetitive tasks frees up time for more strategic analysis and decision-making.
4. **Interactive Analysis:** Pivot tables support interactive data exploration, allowing users to drill down into specific data points or roll up to broader summaries. This interactivity enables users to explore data in a more granular way and adjust their focus as needed. For example, a user can click on a summary figure to view detailed data underlying that summary, providing deeper insights and facilitating a more nuanced understanding of the data.
5. **Error Reduction:** By using built-in aggregation functions and eliminating the need for manual calculations, pivot tables help reduce errors in data analysis. Manual data manipulation is prone to mistakes, such as misapplied formulas or incorrect cell references. Pivot tables automate these processes, minimizing the risk of errors and ensuring more accurate and reliable results. This accuracy is critical for making informed decisions based on the data.
6. **Integration:** Pivot tables can integrate data from multiple sources, making them versatile tools for comprehensive analysis. They can consolidate data from different sheets, files, or databases into a single pivot table. This integration capability allows users to create unified reports and analyses from disparate data sources, facilitating a more holistic view of information and enabling more informed decision-making.

10.16 APPLICATIONS OF PIVOT TABLES

1. **Financial Reporting:** In financial reporting, pivot tables are used to summarize and analyze financial data such as income statements, balance sheets, and cash flow statements. They help in aggregating financial figures by various dimensions like departments, time periods, or account types. This capability enables financial analysts to generate detailed reports, track financial performance, and identify trends or anomalies in financial data.
2. **Sales Analysis:** Pivot tables are extensively used in sales analysis to evaluate sales performance across different dimensions. Users can analyze sales data by product, region, salesperson, or time period to identify high-performing products, regions with the highest sales, or top sales people. This analysis helps businesses understand sales patterns, forecast future sales, and make data-driven decisions to enhance sales strategies.
3. **Customer Analysis:** Businesses use pivot tables to analyze customer data, including purchase behavior, demographics, and feedback. By summarizing data such as customer spending, frequency of purchases, or satisfaction scores, companies can gain insights into customer preferences and behavior. This information is valuable for segmenting customers, tailoring marketing efforts, and improving customer service.
4. **Inventory Management:** Pivot tables are useful for managing inventory by tracking stock levels, categorizing products, and analyzing inventory turnover rates. They help in summarizing data related to stock quantities, reorder levels, and product performance. This analysis assists in optimizing inventory levels, reducing stockouts, and managing inventory more effectively.
5. **Human Resources:** In human resources, pivot tables help analyze employee data, such as headcount, turnover rates, and compensation. They can be used to generate reports on employee demographics, track recruitment metrics, and evaluate workforce distribution. This analysis supports HR decision-making in areas like talent management, compensation planning, and organizational development.
6. **Project Management:** Pivot tables are valuable for project management by summarizing project data, tracking progress, and analyzing resource allocation. They can be used to generate reports on project milestones, budget utilization, and resource availability. This helps project managers monitor project status, assess performance, and make informed decisions to keep projects on track.
7. **Education and Research:** In educational settings, pivot tables are employed to analyze student performance data, track attendance, and generate academic reports. They can summarize data on test scores, grades, and attendance rates to evaluate student progress and identify areas for improvement. Researchers also use pivot tables to organize and analyze experimental data, facilitating a clearer understanding of research findings.
8. **Marketing Analysis:** Pivot tables assist in analyzing marketing campaign results, tracking key performance indicators (KPIs), and assessing the effectiveness of

marketing strategies. By summarizing data such as campaign reach, conversion rates, and ROI, marketers can evaluate the success of their campaigns and adjust strategies to achieve better results.

In summary, pivot tables are a powerful tool for data analysis and reporting, offering significant benefits in terms of summarization, flexibility, efficiency, and interactive exploration. Their versatility and ease of use make them invaluable across various domains, from finance and sales to HR and marketing.

10.17 BUILDING PIVOT TABLE

Excel offers several methods for creating new pivot tables in your worksheets:

- **Quick Analysis Tool:** With all the cells in the data list selected, click the Quick Analysis tool and then select your pivot table on the Tables tab of its drop-down palette.
- **Recommended Pivot Tables Button:** With the cell pointer in one of the cells of a data list, click the Recommended Pivot Tables button on the Insert tab and then select your pivot table in the Recommended Pivot Tables dialog box.
- **Pivot Table Button:** With the cell pointer in one of the cells of a data list, click the Pivot Table button on the Insert tab and then use the Create Pivot Table dialog box to specify the contents and location of your new table before you manually select the fields in the data source to use.

Following sample travelling data will be used to demonstrate the working of pivot table.

1	Month	Types	Category	Amount
2	Jan	Personal	Restaurant	35
3	Jan	Business	Taxi	42
4	Jan	Business	Restaurant	64
5	Jan	Personal	Restaurant	22
6	Jan	Business	Tips	12
7	Feb	Personal	Restaurant	32
8	Feb	Personal	Entertainme	24
9	Feb	Business	Entertainme	45
10	Feb	Business	Lodging	98
11	Mar	Business	Airfare	250
12	Mar	Personal	Restaurant	24
13	Mar	Business	Restaurant	45
14	Mar	Business	Lodging	160
15	Mar	Business	Tolls	8

Data Source: ***Excel 2007 Advanced Report Development, Wiley***

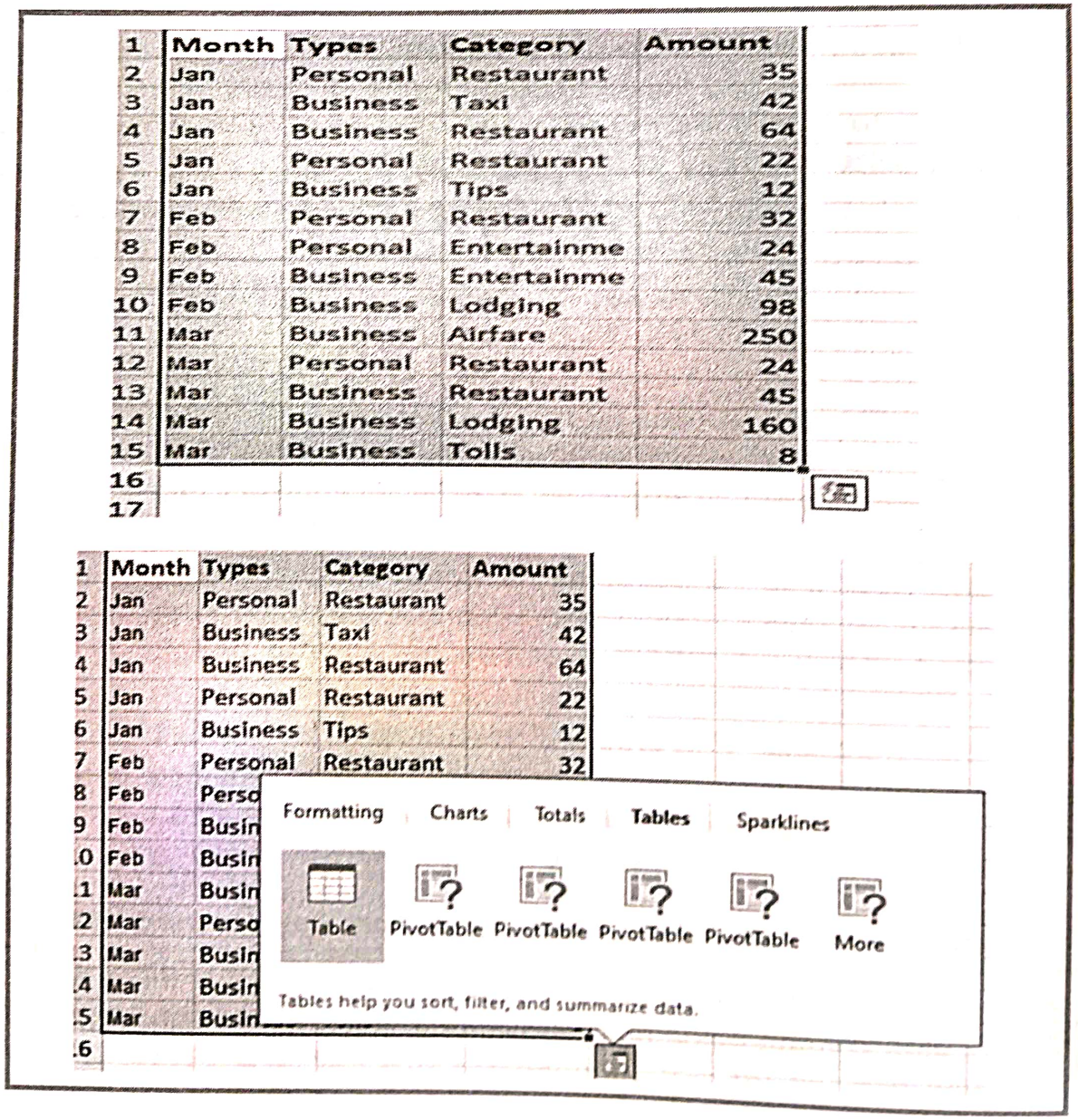
To create the best pivot tables, make sure your data follows these rules:

- Ensure each column has a one-cell heading. Keep the headings unique; don't use the same heading for two columns.

- If a column should contain numeric data, don't allow blank cells in the column. Use zeros instead of blanks.
- Do not use blank rows or blank columns.
- If totals are embedded in your report, remove them.
- If you add new data to the bottom of your data set each month, you should strongly consider converting your data set to a table. Pivot tables created from tables automatically pick up new rows pasted to the bottom of the tables after a refresh.

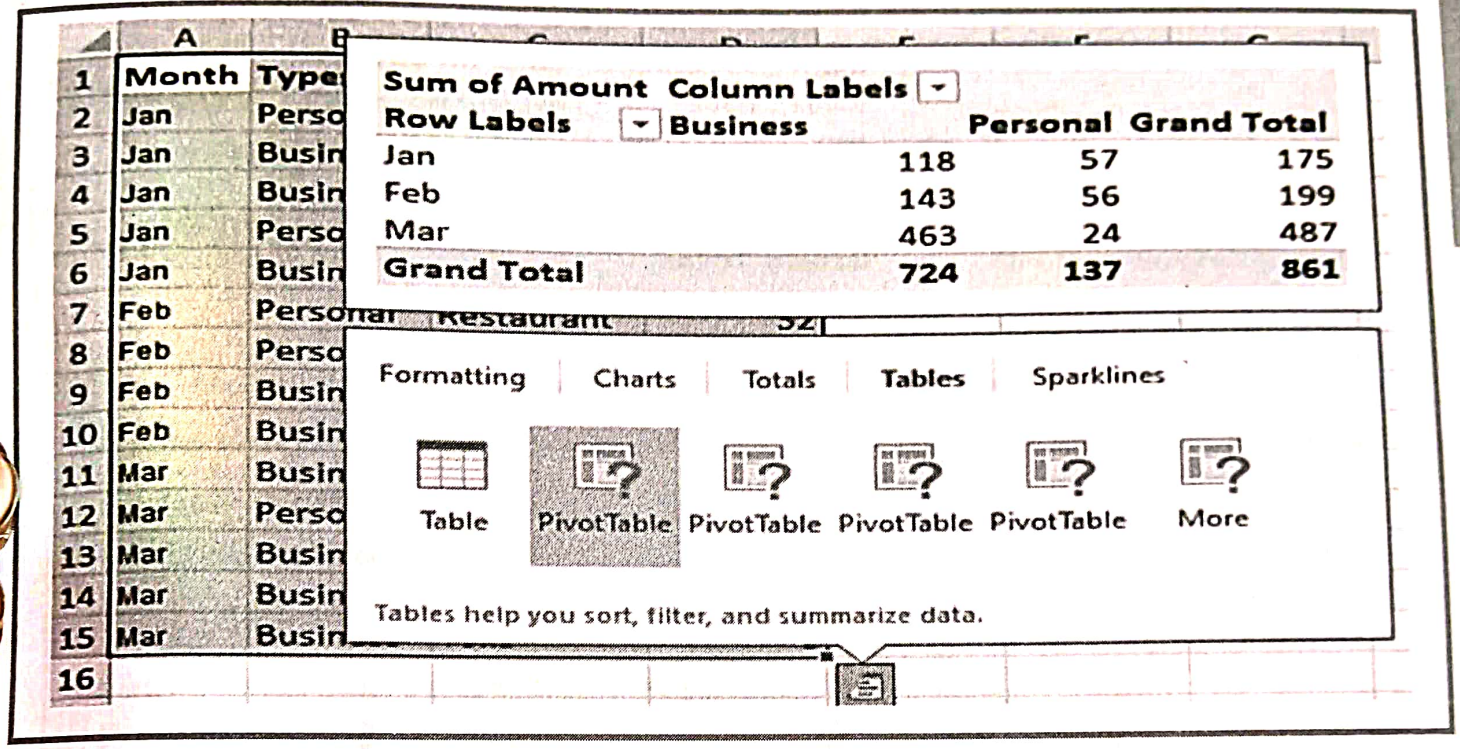
10.17.1 CREATING PIVOT TABLE USING QUICK ANALYSIS TOOL

1. Select all the data (including the column headings) in your data list as a cell range in the worksheet. Click the Quick Analysis tool that appears right below the lower-right corner of the current cell selection.
2. Click the Tables tab at the top of the Quick Analysis options palette. To preview each pivot table that Excel 2016 can create for your data, highlight its Pivot Table button in the Quick Analysis palette.



	Month	Types	Category	Amount
2	Jan	Personal	Restaurant	35
3	Jan	Business	Taxi	42
4	Jan	Business	Restaurant	64
5	Jan	Personal	Restaurant	22
6	Jan	Business	Tips	12
7	Feb	Personal	Restaurant	32
8	Feb	Personal	Entertainme	24
9	Feb	Business	Entertainme	45
10	Feb	Business	Lodging	98
11	Mar	Business	Airfare	250
12	Mar	Personal	Restaurant	24
13	Mar	Business	Restaurant	45
14	Mar	Business	Lodging	160
15	Mar	Business	Tolls	8

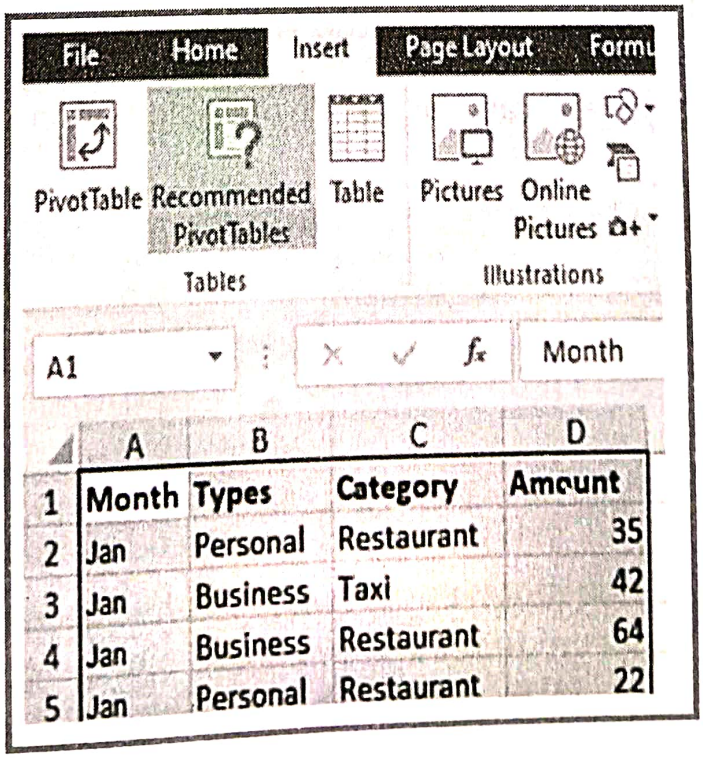
3. When a preview of the pivot table you want to create appears, click its button in the Quick Analysis options palette to create it.

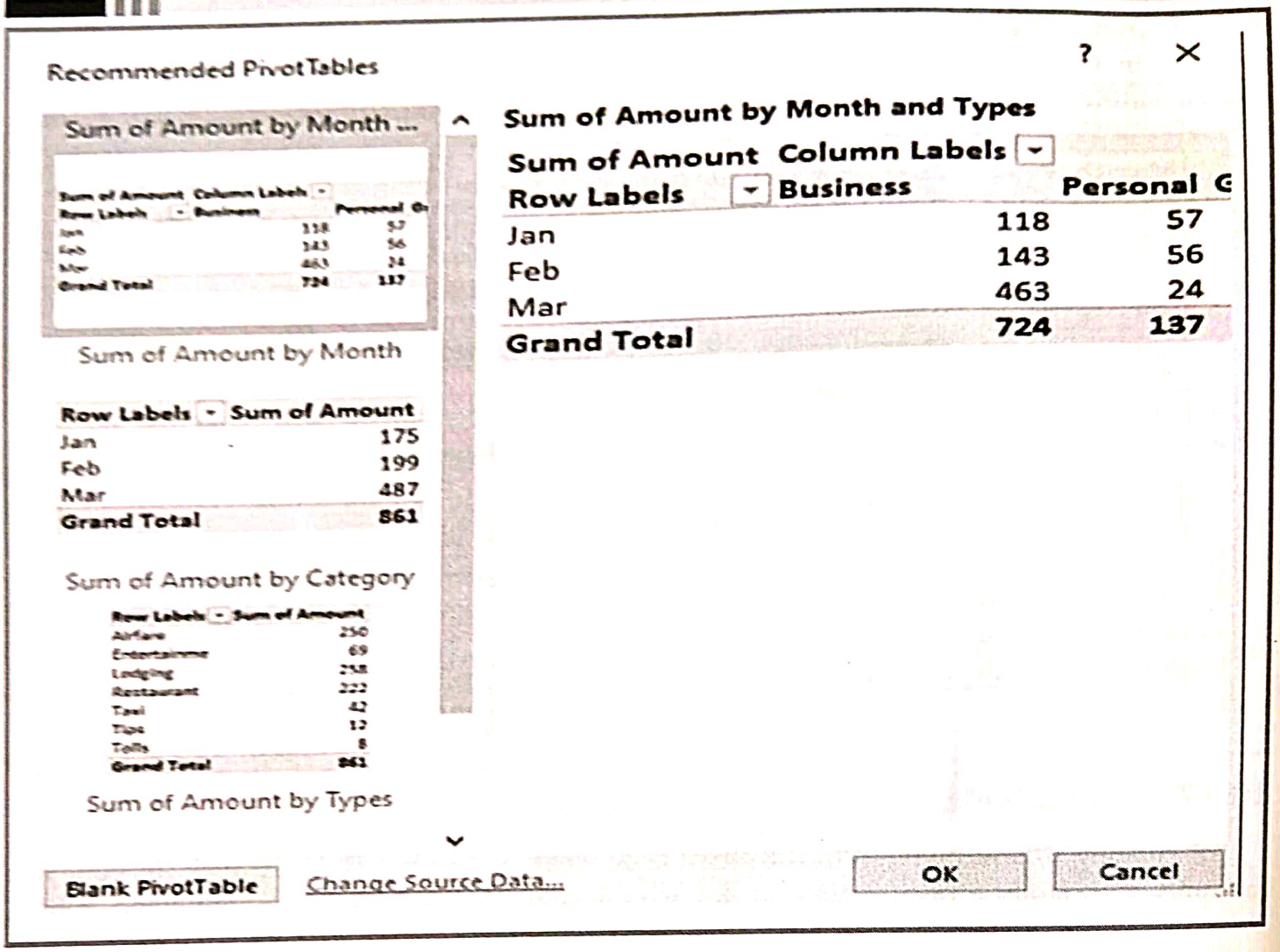


10.17.2 RECOMMENDED PIVOT TABLES

If creating a new pivot table with the Quick Analysis tool is too much work for you, you can quickly generate a pivot table with the new Recommended Pivot Tables command button.

1. Select a cell in the data list for which you want to create the new pivot table.
2. Click the Recommended Pivot Tables command button on Insert tab of the Ribbon.
3. Select the sample of the pivot table you want to create in the list box on the left and then select OK.

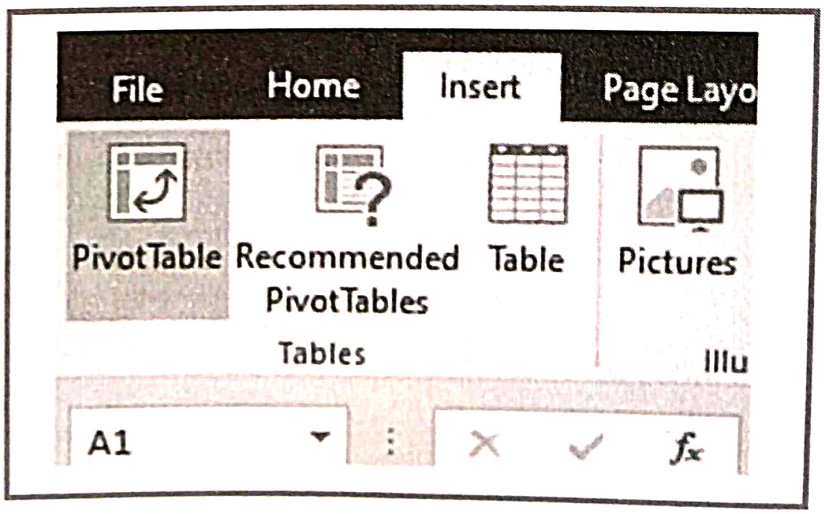


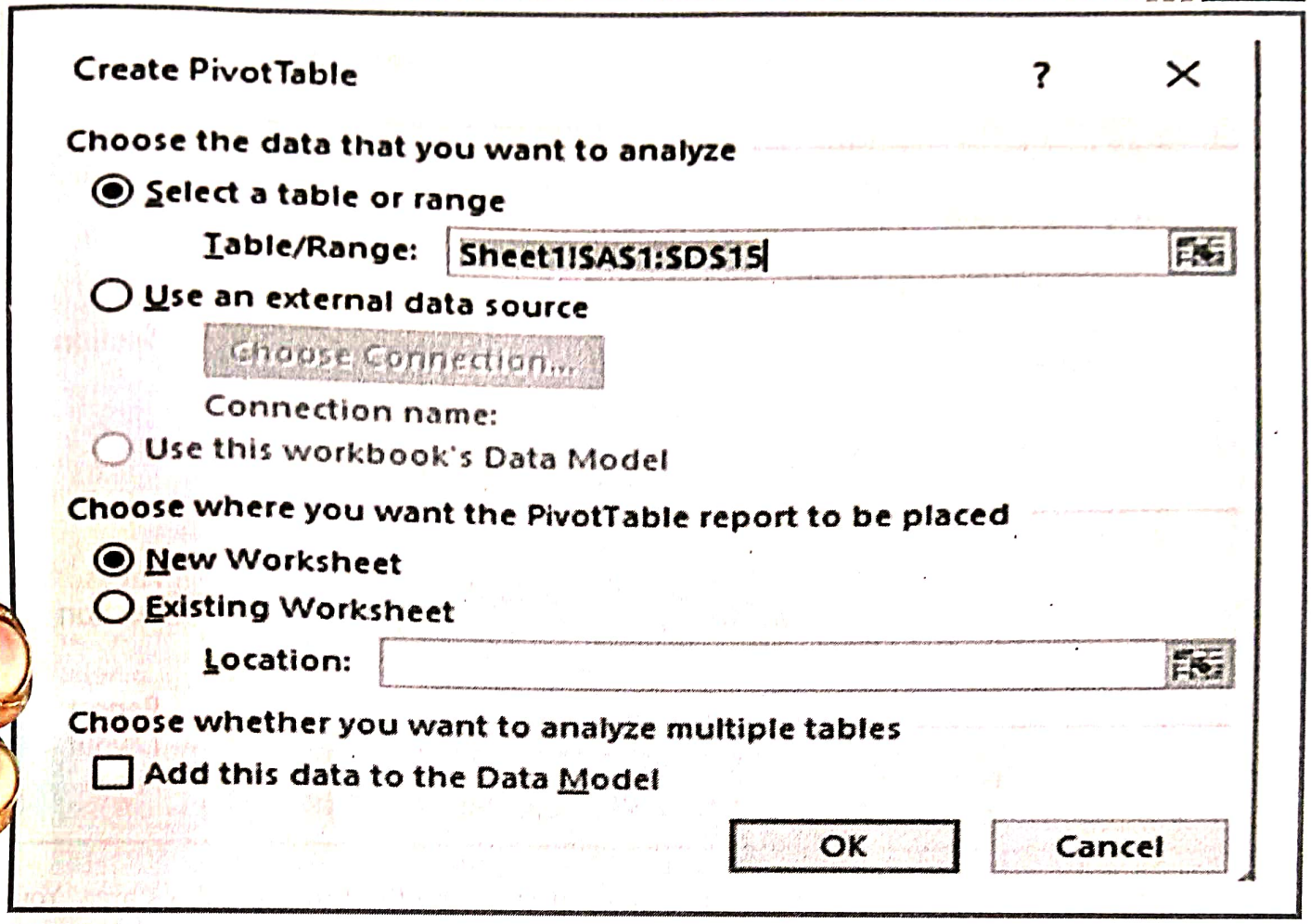


10.17.3 PIVOT TABLE BUTTON/CREATING NEW PIVOT TABLE

When you want your pivot table to work with data from fields in more than one (related) data table or with data stored in a data table that doesn't reside in your worksheet as when connecting with an external data source you need to manually create the pivot table.

The process of creating a Pivot Table is pretty simple. It all starts on the Insert tab, as shown in Figure where you can see the Pivot Table button highlighted. A quick click and the dialog



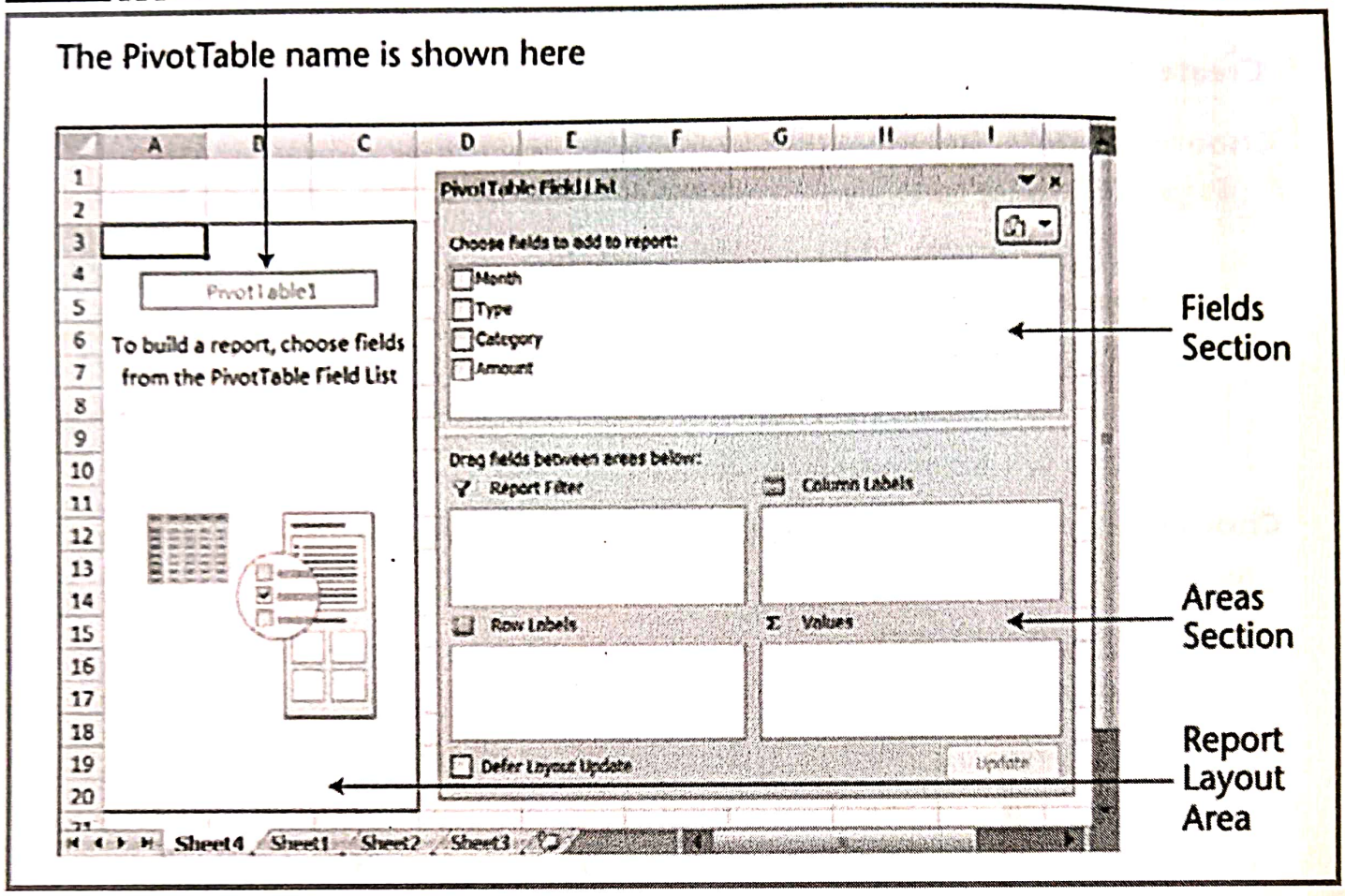


One of the following data source can be used to create pivot table.

1. **Select a table or range**: This would normally be a worksheet that you already have open, which as shown in Figure automatically includes all the records in the active sheet.
2. **Use an external data source.** This option opens an Existing Connections dialog box, through which you can tap into linked data from the open workbook.
3. **Choose where you want the Pivot Table report to be placed**: Your choices here are New Worksheet (this is the default) or Existing Worksheet.

After selection new worksheet following figure is shown (next page)

- **Filters:** Filters allow you to display just the data you want to see and the rest of the data will be hidden. When you move a field to the Filters area you can then choose which items in the field you want to see or not see by checking or unchecking the boxes. They work in the same way as a filter in a table.
- **Rows:** Rows show each item in the field in separate rows. One field item occupies one row. They go down the left side of the Pivot Table.
- **Columns:** Columns shows each item in the field across separate columns.
- **Values:** Values shows the summary information. Excel offers many ways to summarizes your information which include SUM, COUNT, AVERAGE, MAX,



MIN and so on. You are required to enter at least one field in the Values area. You can have the same field in the Values area twice.

10.18 DIFFERENT LEVELS TO PREPARE THE PIVOT TABLE

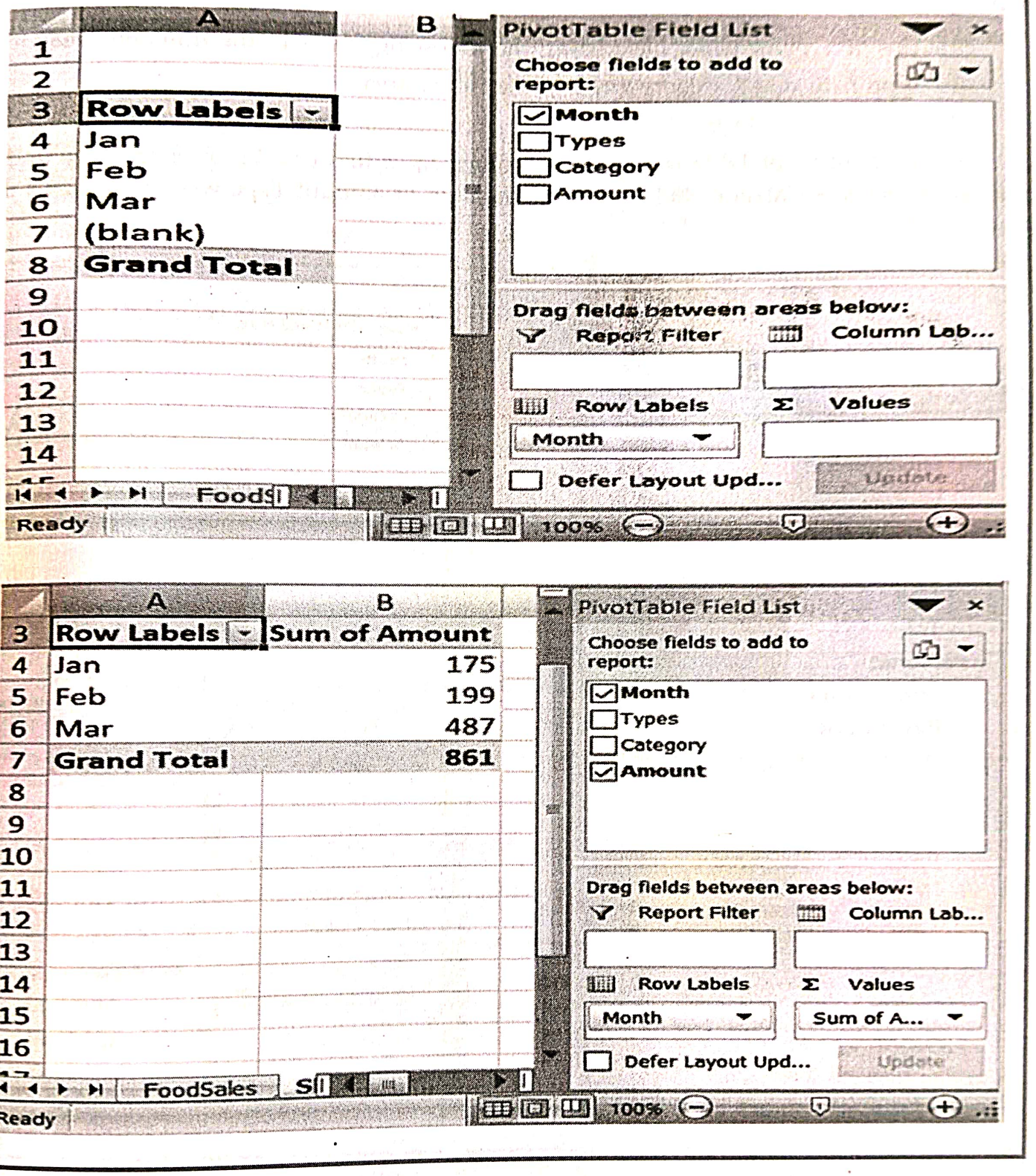
Next, set up the actual layout of the pivot table. You can do so by using of the following techniques:

- Drag the field names (at the top of the Pivot Table Fields task pane) to one of the four boxes at the bottom of the task pane.
- Place a check mark next to the item at the top of the Pivot Table Fields task pane. Excel places the field into one of the four boxes at the bottom. You can drag it to a different box, if necessary.
- Right-click a field name at the top of the Pivot Table Fields task pane and choose its location from the shortcut menu (for example, Add to Row Labels).

10.18.1 SINGLE LEVEL PIVOT TABLE

To add the desired field from the sample data in pivot table follows the steps below:

1. Click on the month field and add in the row area.
2. Now click on Amount field and add it to the value area. It will sum the data month wise and then calculates the grand total of all the months.



10.18.2 MULTILEVEL PIVOT LEVEL

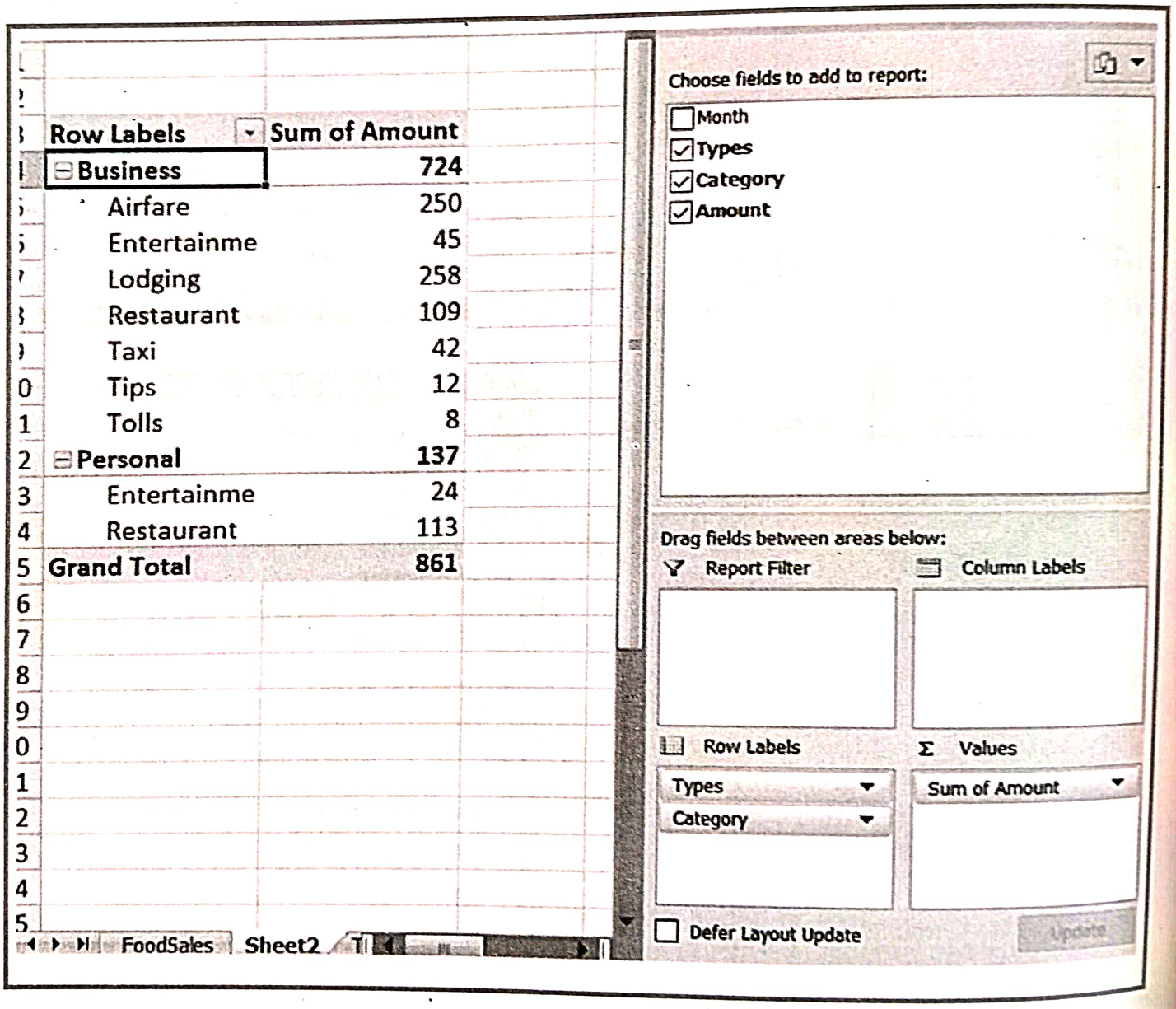
It's perfectly ok to drag more than one field to an area in a pivot table. We will look at an example of **multiple row fields, multiple value fields** and **multiple report filter fields**

Multiple Row Fields

First, insert a pivot table. Next, drag the following fields to the different areas.

1. Type field and Category field to the Rows area.
2. Amount field to the Values area.

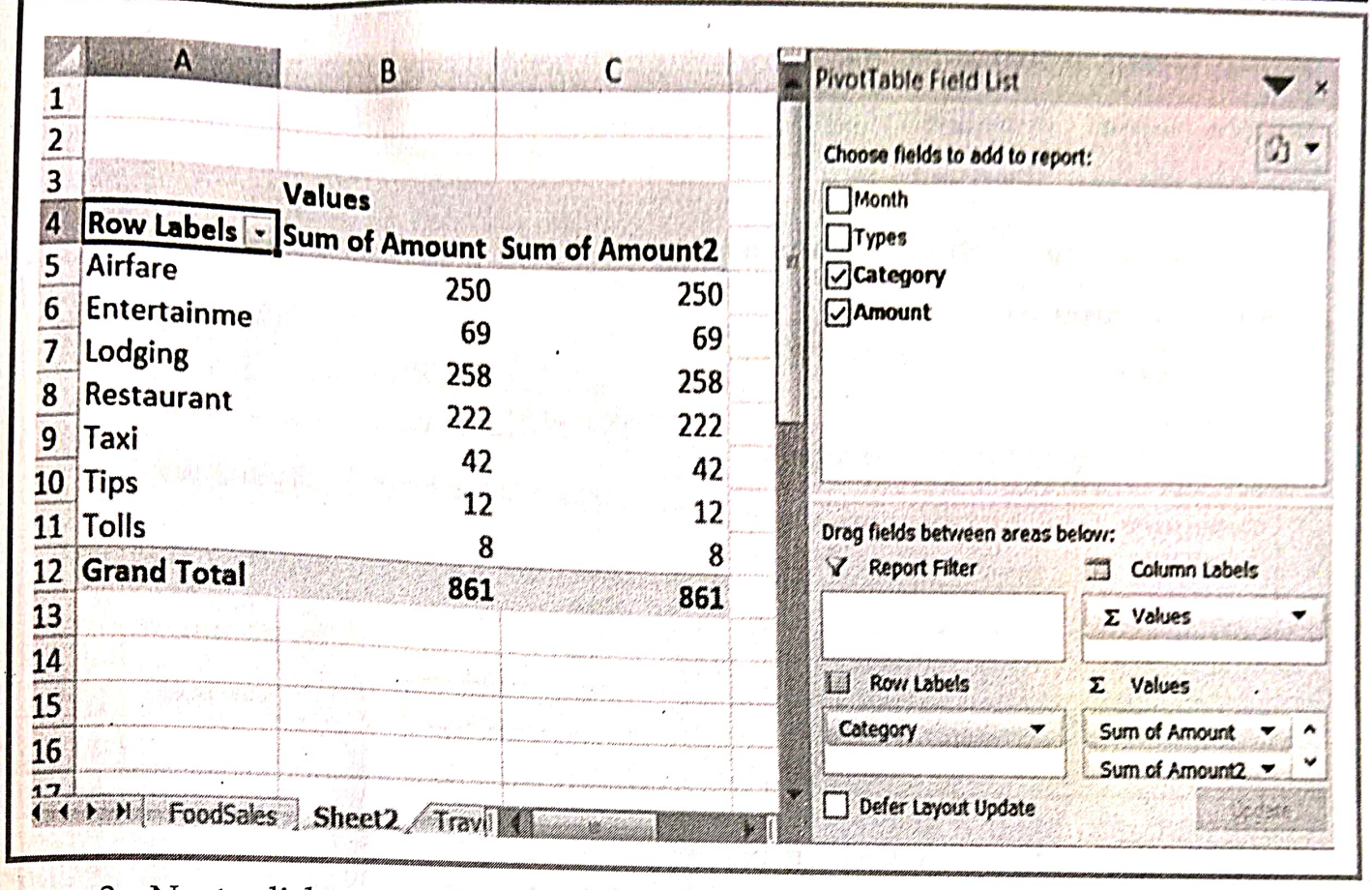
The resultant Pivot Table is show in figure in which first Level is Type field the under that second level is Category field. It shows the sum of amount Type wise then shows the grand total.



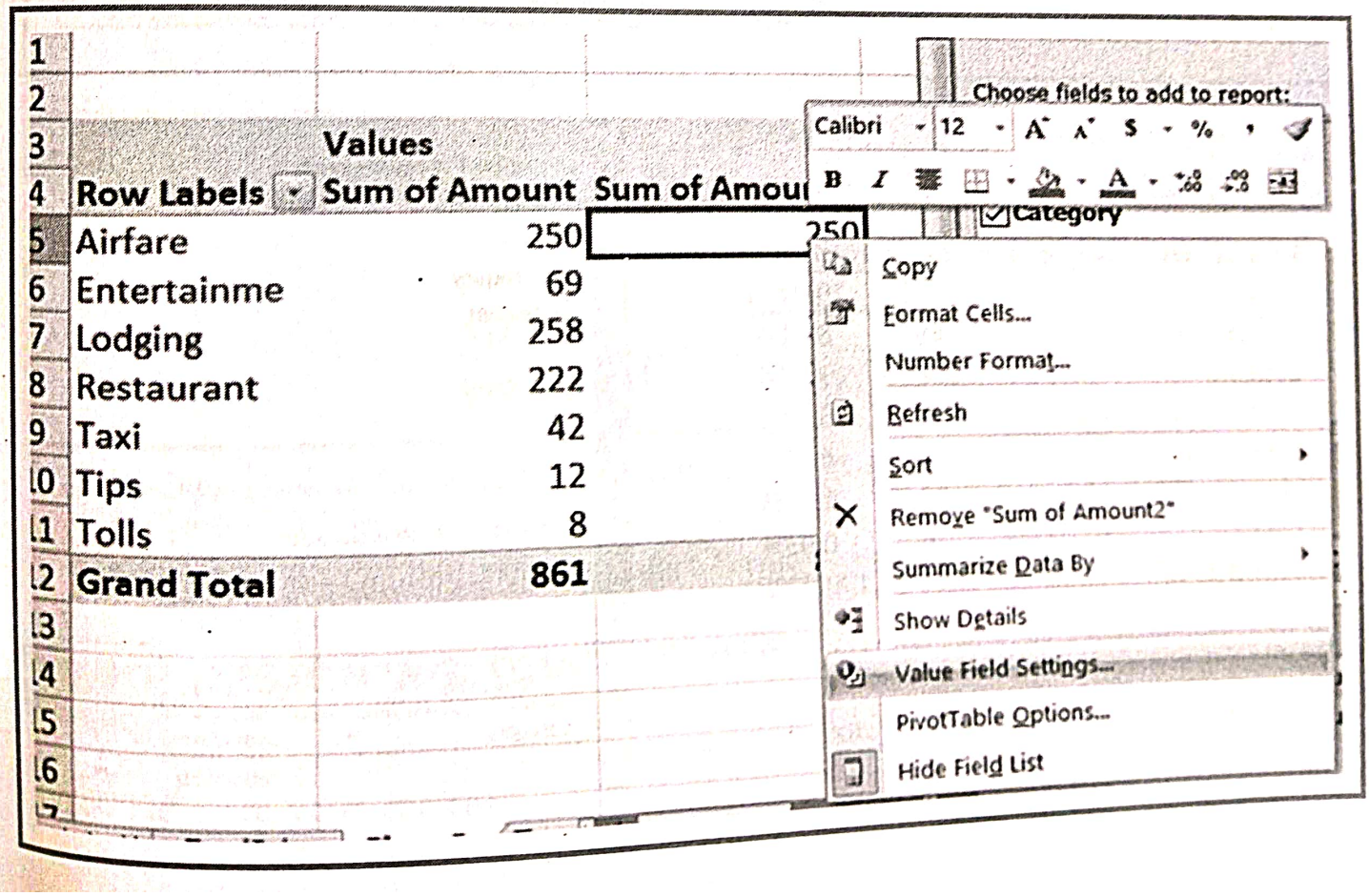
Multiple Value Fields

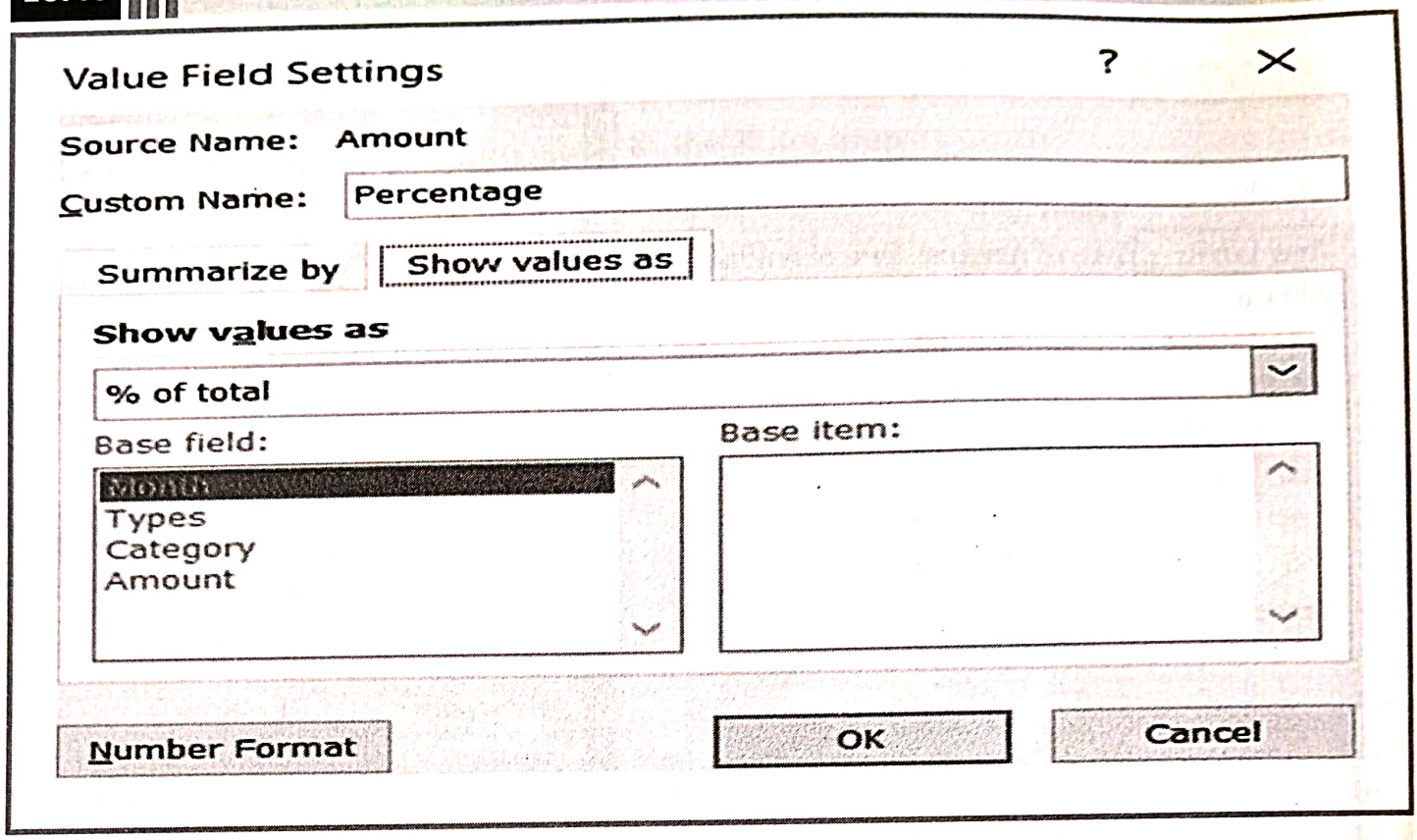
First, insert a pivot table. Next, drag the following fields to the different areas.

1. Country field to the Rows area.
2. Amount field to the Values area (2x).



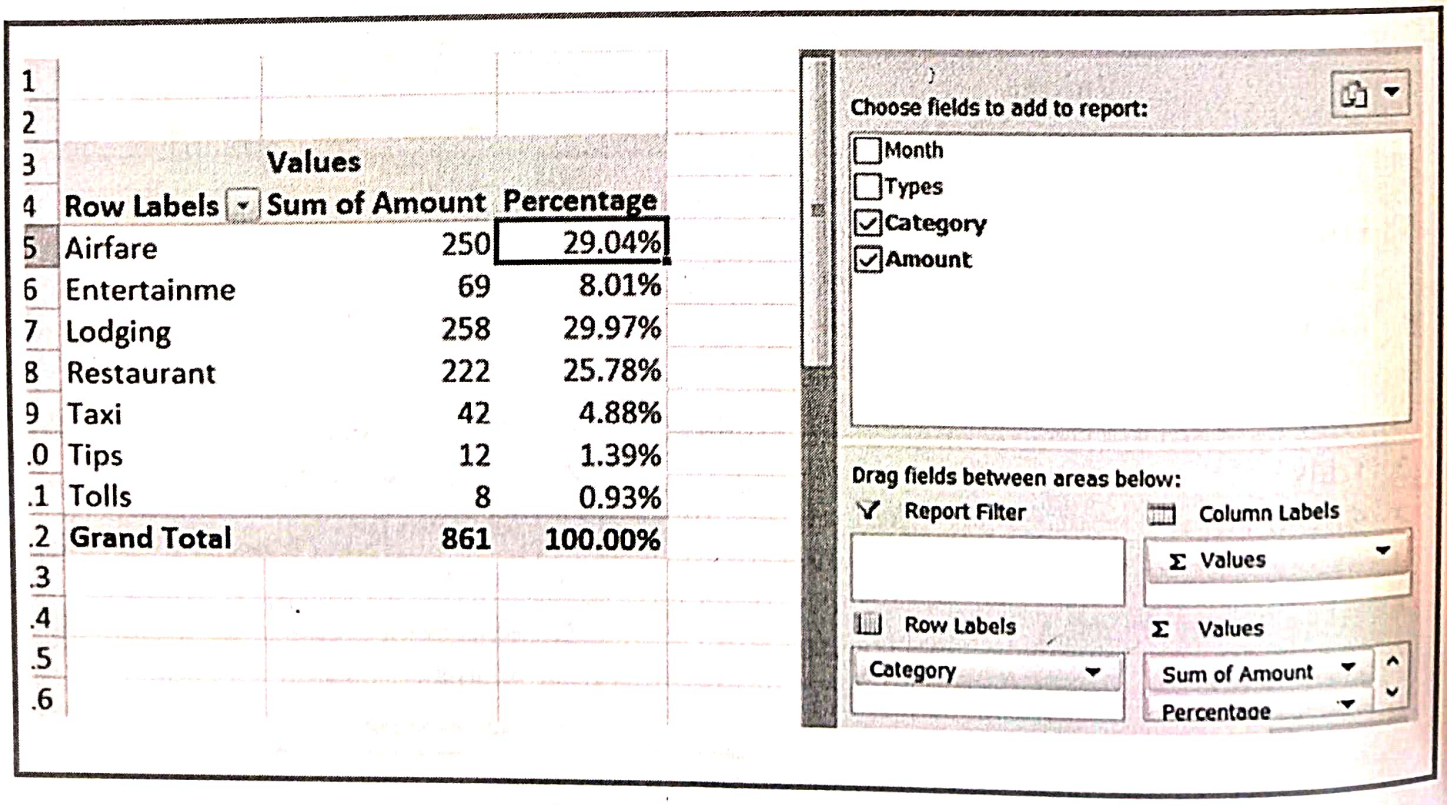
3. Next, click any cell inside the Sum of Amount 2 column.
4. Right click and click on Value Field Settings.





5. Enter Percentage for Custom Name.
6. On the Show Values As tab, select % of Total.
7. Click OK.

The resultant figure is shown below:

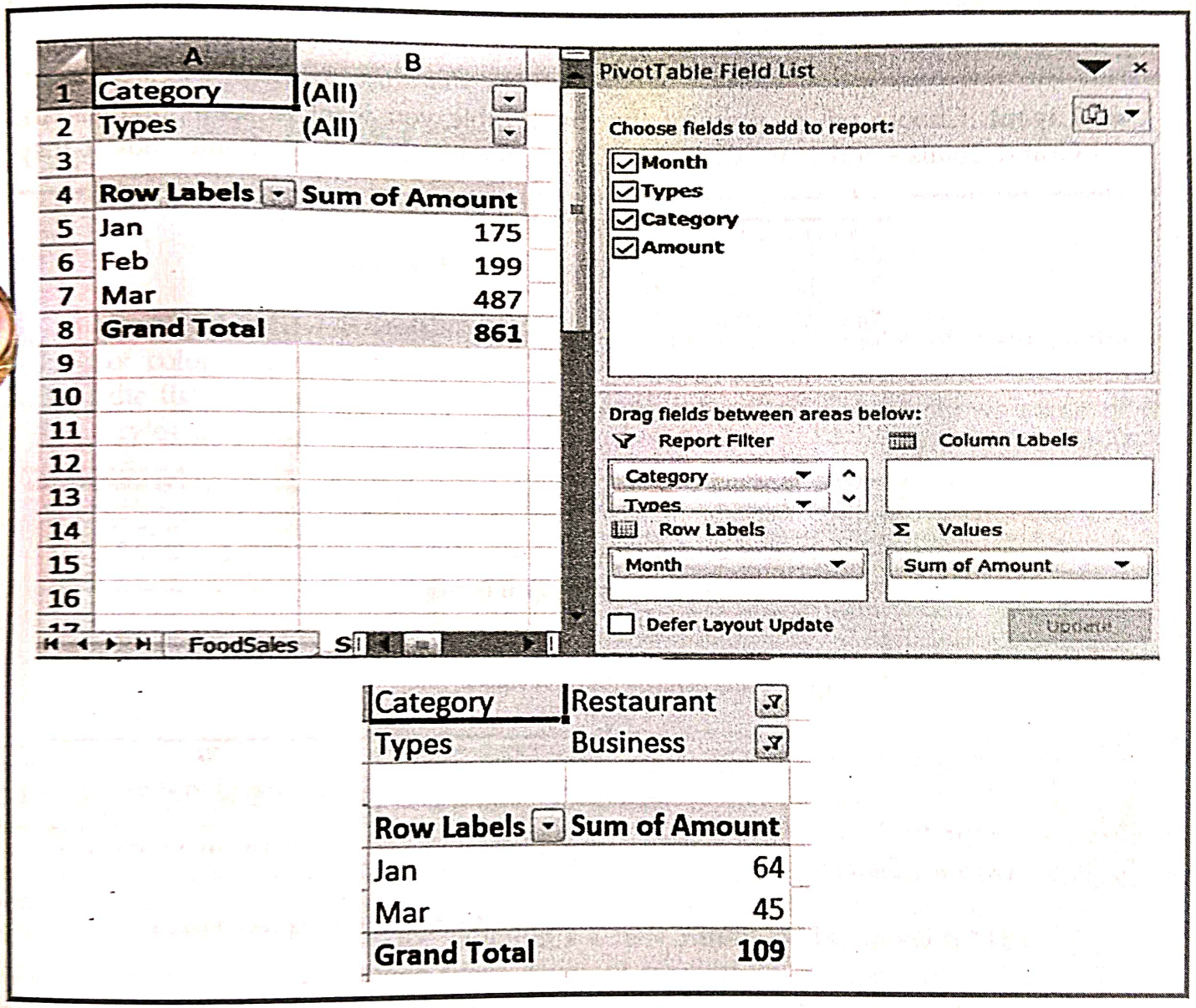


	Values	
Row Labels	Sum of Amount	Percentage
Airfare	250	29.04%
Entertainme	69	8.01%
Lodging	258	29.97%
Restaurant	222	25.78%
Taxi	42	4.88%
Tips	12	1.39%
Tolls	8	0.93%
Grand Total	861	100.00%

Multiple Report Filter Fields

First, insert a pivot table. Next, drag the following fields to the different areas.

1. Category and Type to the Report Filter area.
2. Amount field to the Values area.
3. Month field to the row area.

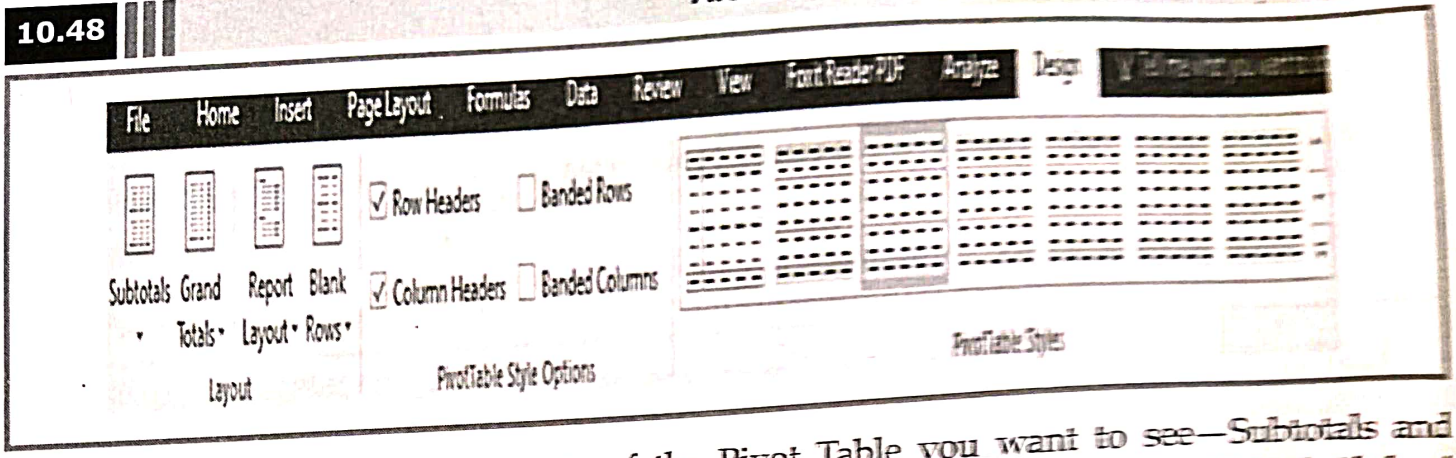


4. Next, select Restaurant from the first filter drop-down and Business from the second filter drop-down.

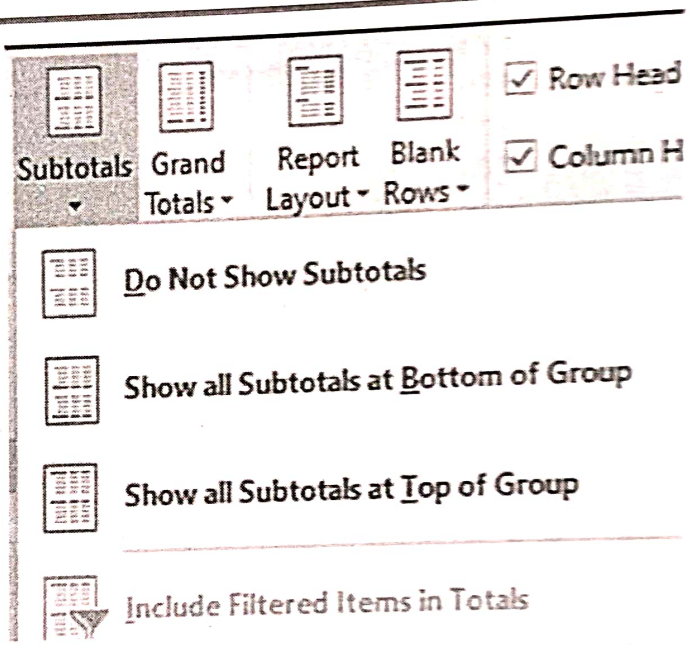
The pivot table shows all the Restart Business in different months.

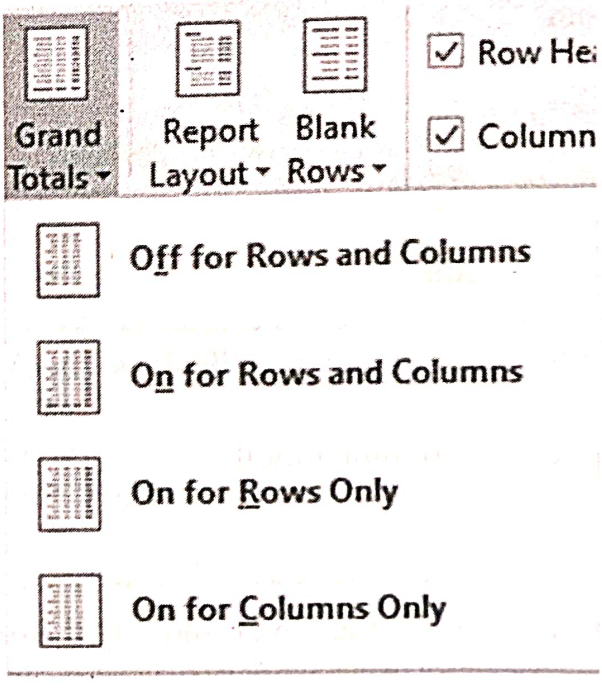
10.19 DESIGNING PIVOT TABLE

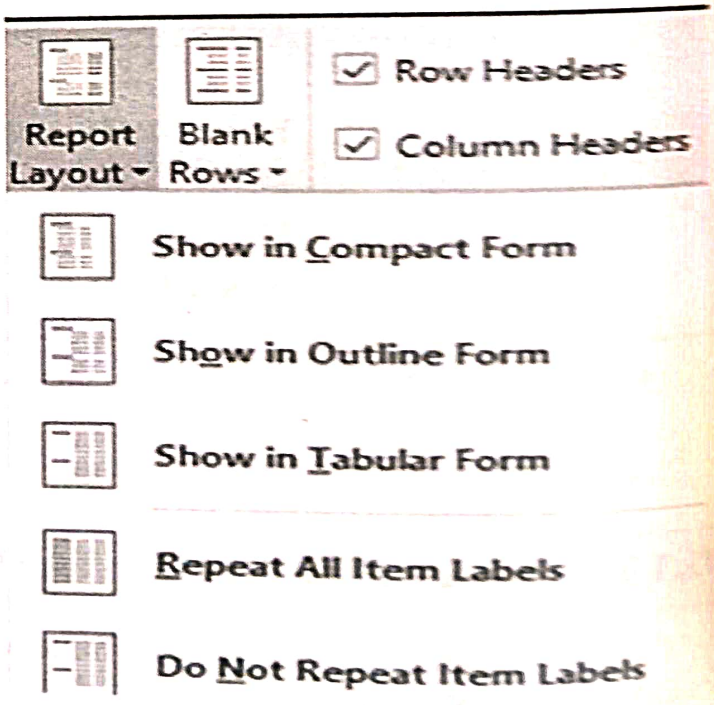
On the Design tab you can determine what you see in the Pivot Table and how it looks. Here are your choices.



- **Layout.** Choose which parts of the Pivot Table you want to see—Subtotals and Grand Totals—and from their drop-down menus, how they should look if they'

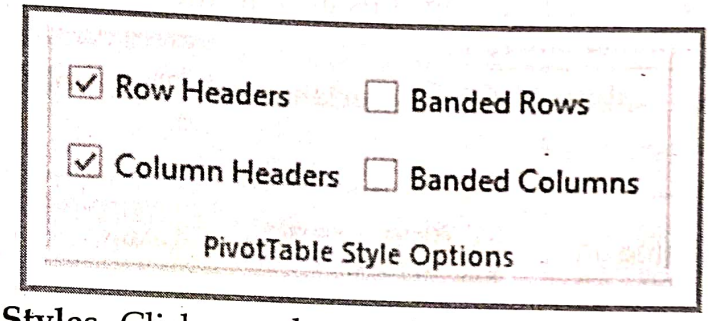




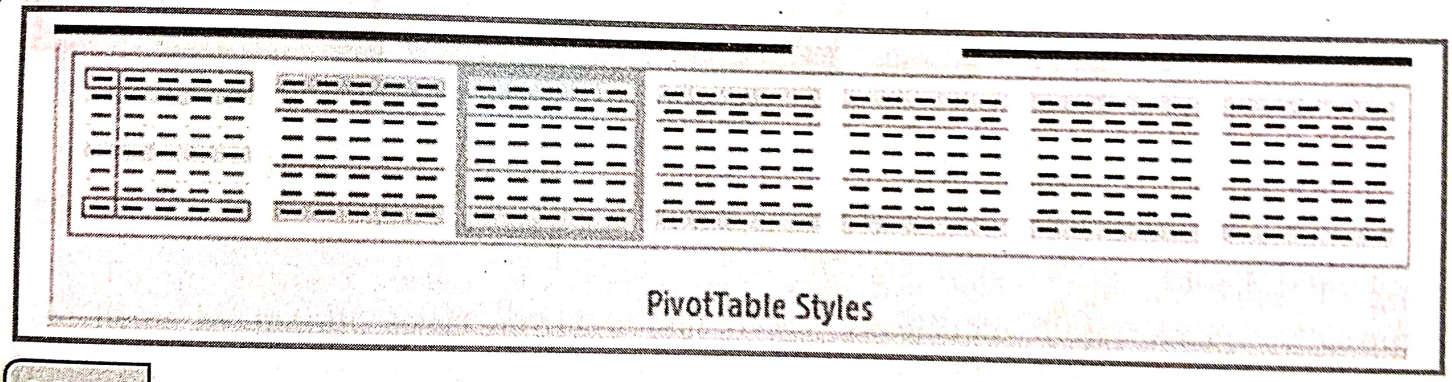


redisplayed. The Report Layout option allows you to show your PivotTable in Compact, Outline, or Tabular format, and to choose how your item labels appear.

- **Pivot Table Style Options.** Here, use the check boxes to decide which things are included—Row Headers, Column Headers(both on by default), and whether or not to use Banded Rows and Columns. If you turn one or both of these on, the options in the next bullet point, PivotTable Styles, change to show banded options. Bands are colored rows and/or columns to help give your PivotTable a more graphical, colorful look.



- **Pivot Table Styles.** Click on a button for any style that looks interesting—the use of colors, number of bands (colored rows/columns), and so on. Scroll through the list of buttons and test any that look interesting. Figure shows some of the styles in place.



10.20 FORMATTING PIVOT TABLE

Even after applying a table style to your new pivot table, you may still want to make some individual adjustments to its formatting, such as selecting a new font, font size, or cell alignment for the text of the table and a new number format for the values in the table's data cells. You can make these types of formatting changes to a pivot table by selecting the part of the table to which the formatting is to be applied and then selecting the new formatting from the appropriate command buttons in the Font, Alignment, and Number groups on the Home tab of the Ribbon.

- **The Home tab** offers formatting options for changing the fonts, font sizes, colors, and fills applied to the cells in your PivotTable The PivotTable Tools section of the
- **Ribbon** (activated whenever you're in a PivotTable) offers an Options and a Design tab. The Design tab (discussed in pervious section), offers layout options for choosing what you see in your PivotTable and how the rows and columns are structured.

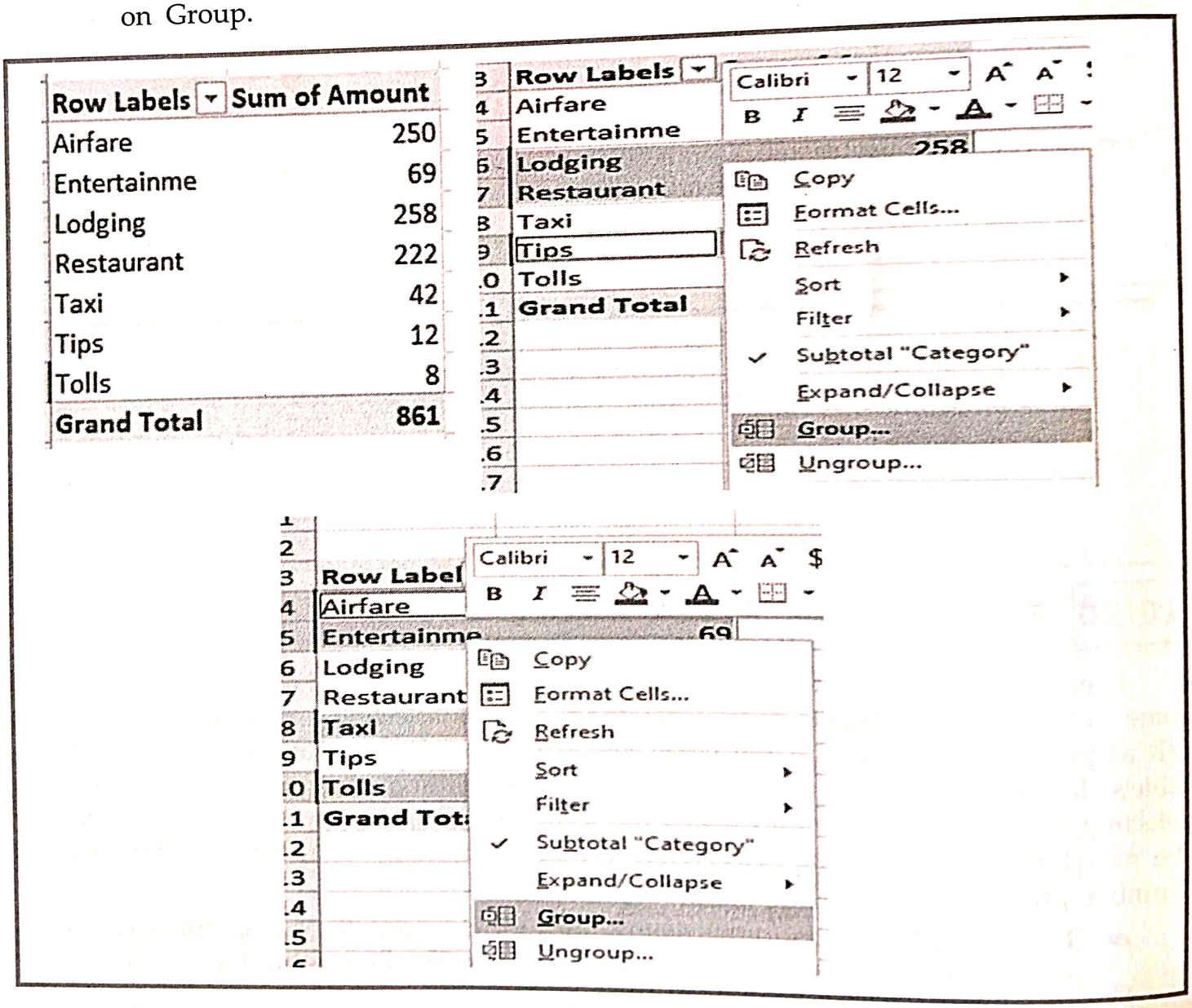
10.21 GROUPING FILEDS IN PIVOT TABLE

This example teaches you how to group pivot table items. Learn how to group products and how to group dates.

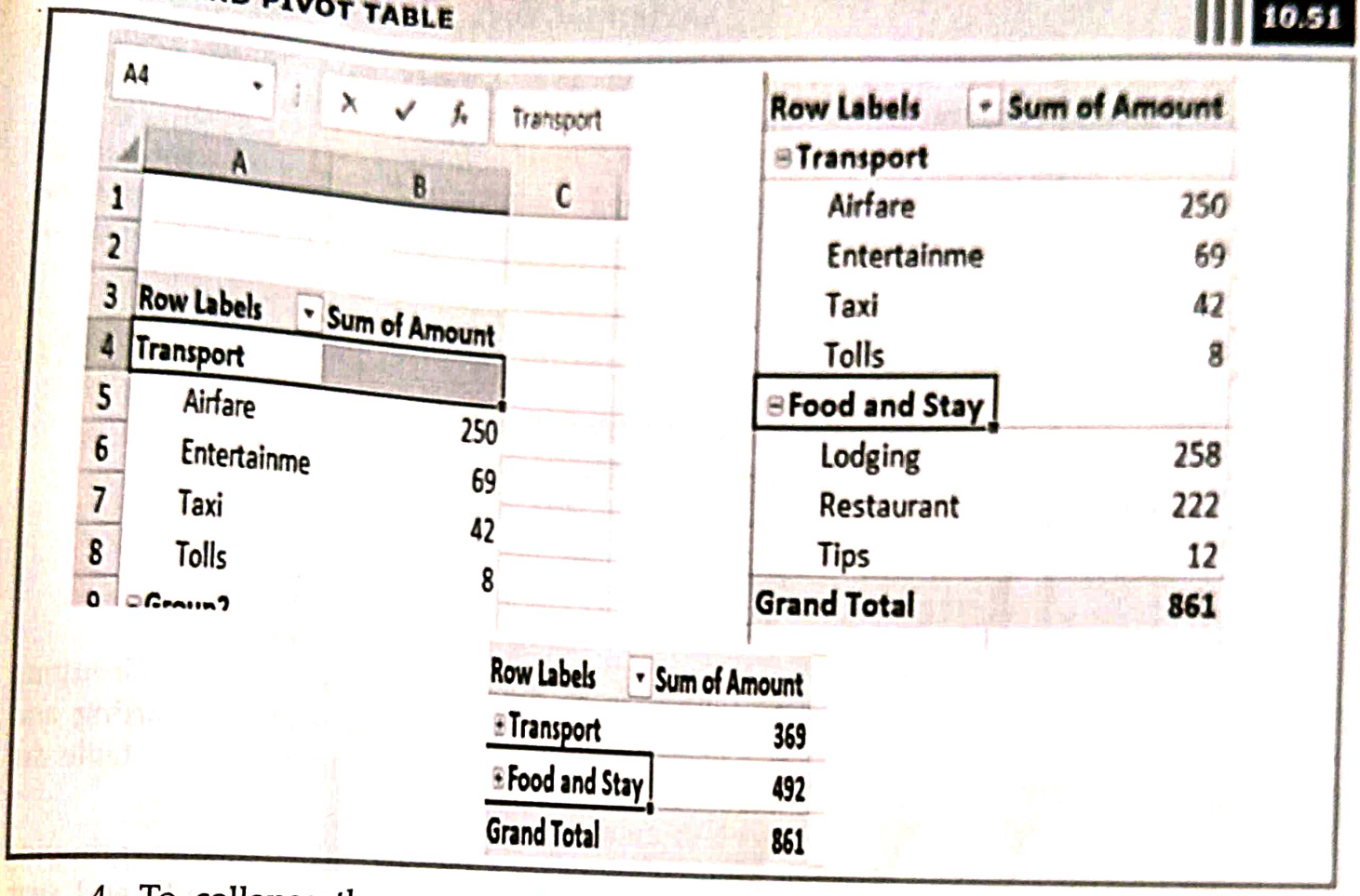
Group Categories

The Category field contains 7 items. To create two groups, execute the following steps.

1. In the pivot table, select Lodging, Restaurant and Tips and Right click and click on Group.
2. In the pivot table, select Airfare, Entertainment, Taxi, Tolls. Right click and click on Group.



3. To change the name of a group (Group 1 or Group 2), select the name, and edit the name in the formula bar. To ungroup, select the group, right click and click on Ungroup.

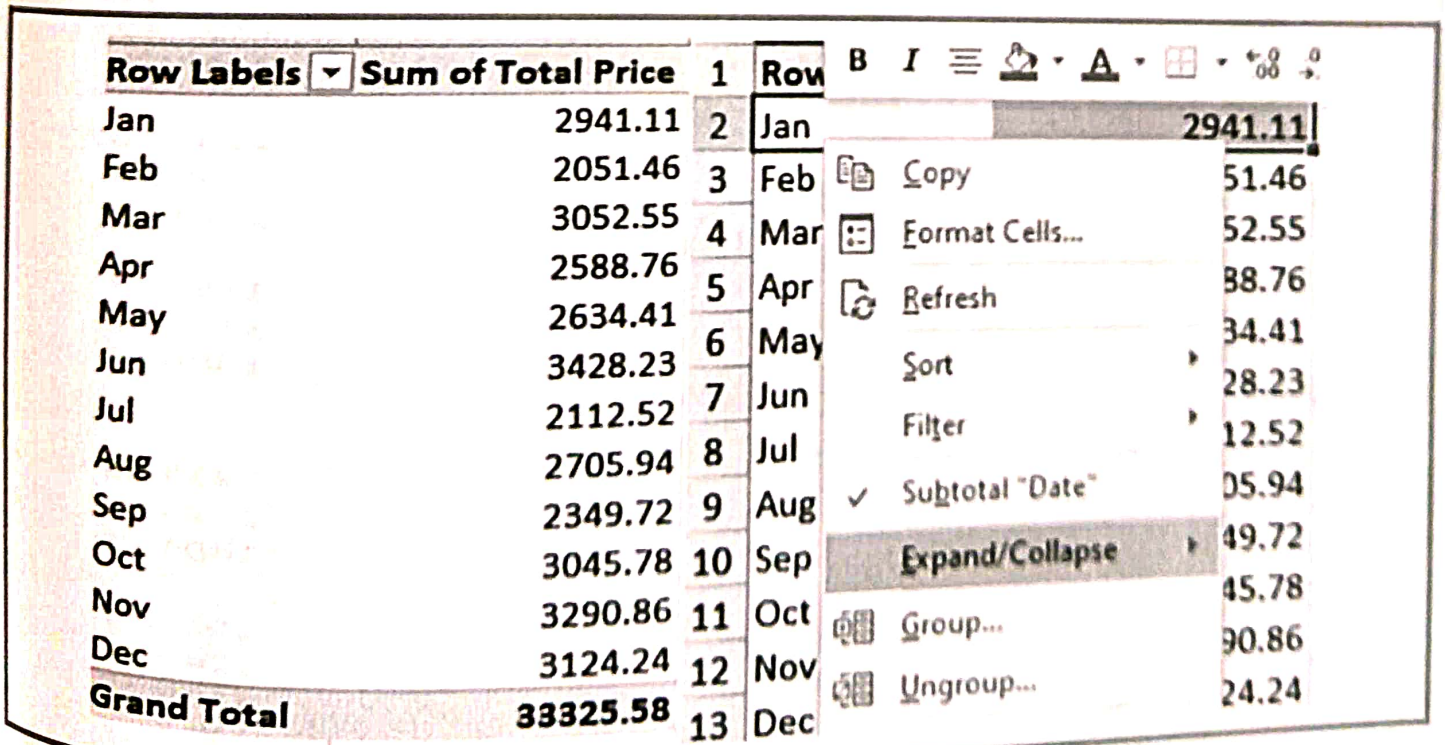


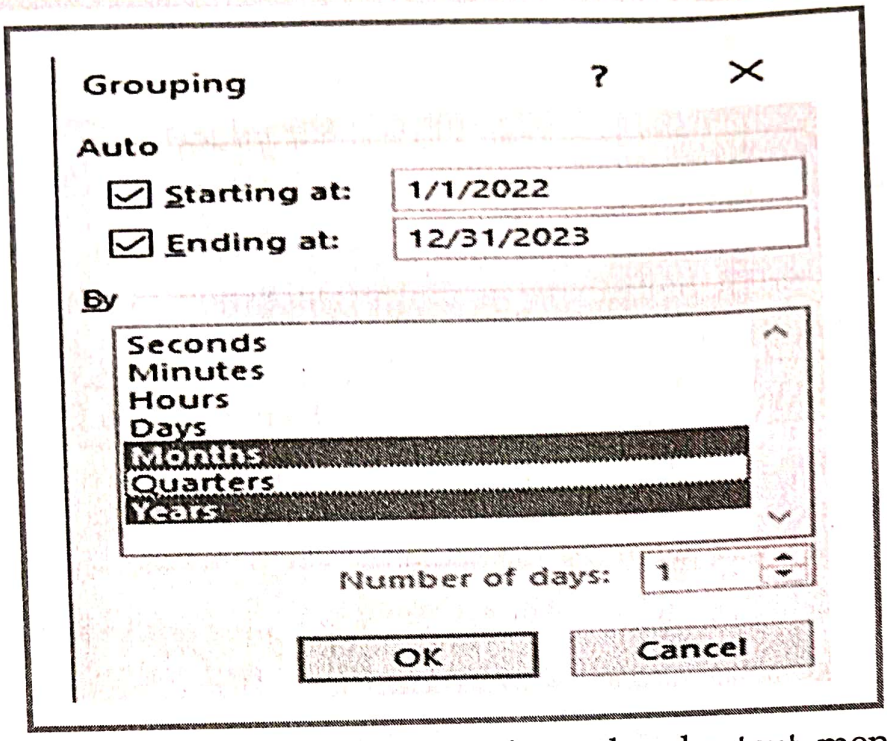
4. To collapse the groups, click the minus signs.

Last figure shows that Food and Stay expenses are higher than Transport.

Group Dates

Figure shows a portion of a simple table with two fields: Date and Sales. To group the items by month, select any date and choose PivotTable Tools ➪ Analyze ➪ Group ➪





Group Field (or right-click and choose Group from the shortcut menu). The Grouping dialog box. In the By list box, select Months and Years and verify that the starting and ending dates are correct for your data. Click OK. The Date items in the pivot table are grouped by years and by months

Following two figures tow different views of grouping dates. First figure shows grouping of month and year and second one shows grouping of quarter along with month and year.

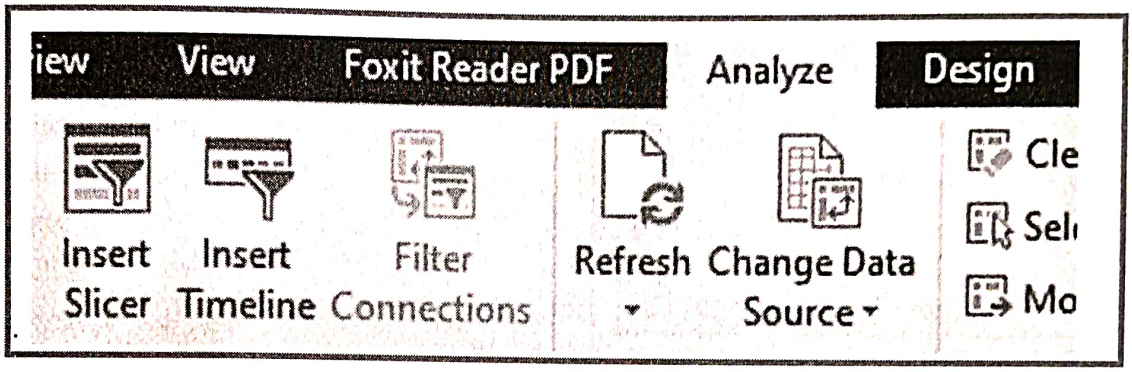
Row Labels	Sum of Total Price
⊟2022	
Jan	1705.82
Feb	926.05
Mar	1647.7
Apr	1052.09
May	1393.45
Jun	2308.75
Jul	1263.16
Aug	1490.9
Sep	1524.06
Oct	1755.57
Nov	1311.47
Dec	1609.64
⊟2023	
Jan	1235.29
Feb	1125.41
Mar	1404.85
Apr	1536.67
May	1240.96
Jun	1119.48

⊟Qtr1	
Jan	1705.82
Feb	926.05
Mar	1647.7
⊟Qtr2	
Apr	1052.09
May	1393.45
Jun	2308.75
⊟Qtr3	
Jul	1263.16
Aug	1490.9
Sep	1524.06
⊟Qtr4	
Oct	1755.57
Nov	1311.47
Dec	1609.64
2023	
⊟Qtr1	
Jan	1235.29
Feb	1125.41
Mar	1404.85
⊟Qtr2	
Apr	1536.67
May	1240.96
Jun	1119.48

10.22 SLICING DATA IN PIVOT TABLE

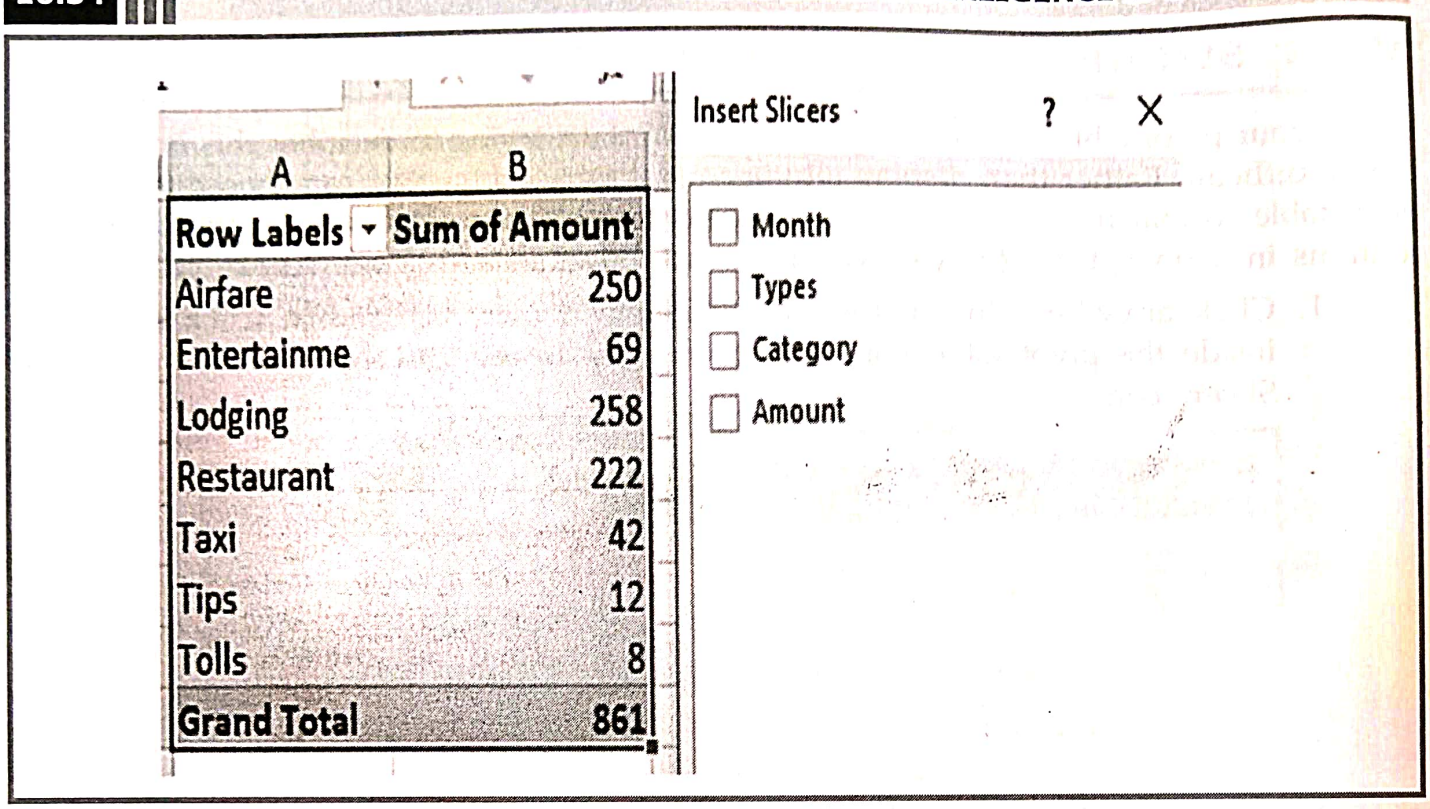
If your pivot table contains large amounts of data, trying to decipher this information can be difficult. Rather than display all your pivot table's data, you can choose to slice the pivot table so that it displays only the specific data you want to view. To turn rows into columns in a pivot table (or vice versa), follow these step.

1. Click anywhere inside the pivot table you want to modify. Click anywhere inside the pivot table you want to modify. In the Filter group, click the Insert Slicer icon.



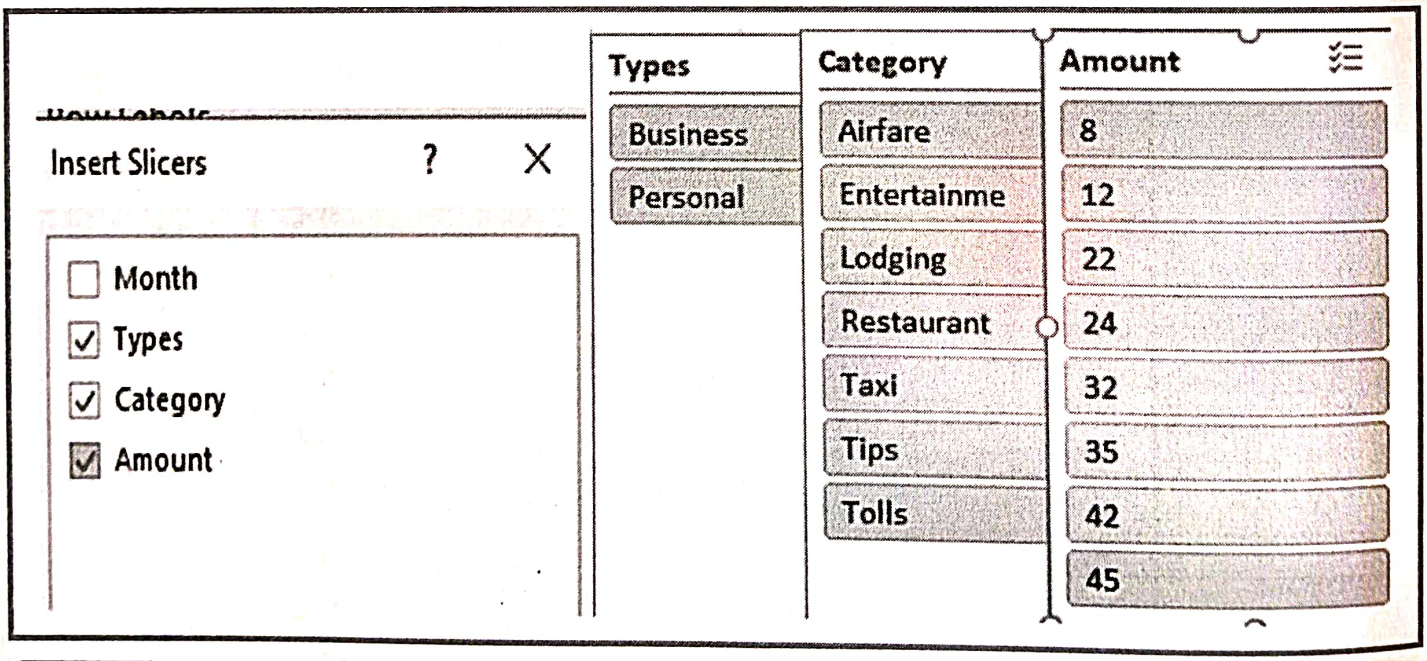
2. In the Insert Slicers dialog box, select one or more check boxes and then click OK. To move a slicer pane to a more convenient location on the screen, click the pane to select it, and then move the mouse pointer over the slicer pane border and drag the mouse.

Month	Types	Category	Amount
Jan	Personal	Restaurant	35
Jan	Business	Taxi	42
Jan	Business	Restaurant	64
Jan	Personal	Restaurant	22
Jan	Business	Tips	12
Feb	Personal	Restaurant	32
Feb	Personal	Entertainme	24
Feb	Business	Entertainme	45
Feb	Business	Lodging	98
Mar	Business	Airfare	250
Mar	Personal	Restaurant	24
Mar	Business	Restaurant	45
Mar	Business	Lodging	160
Mar	Business	Tolls	8



Row Labels	Sum of Amount
Airfare	250
Entertainme	69
Lodging	258
Restaurant	222
Taxi	42
Tips	12
Tolls	8
Grand Total	861

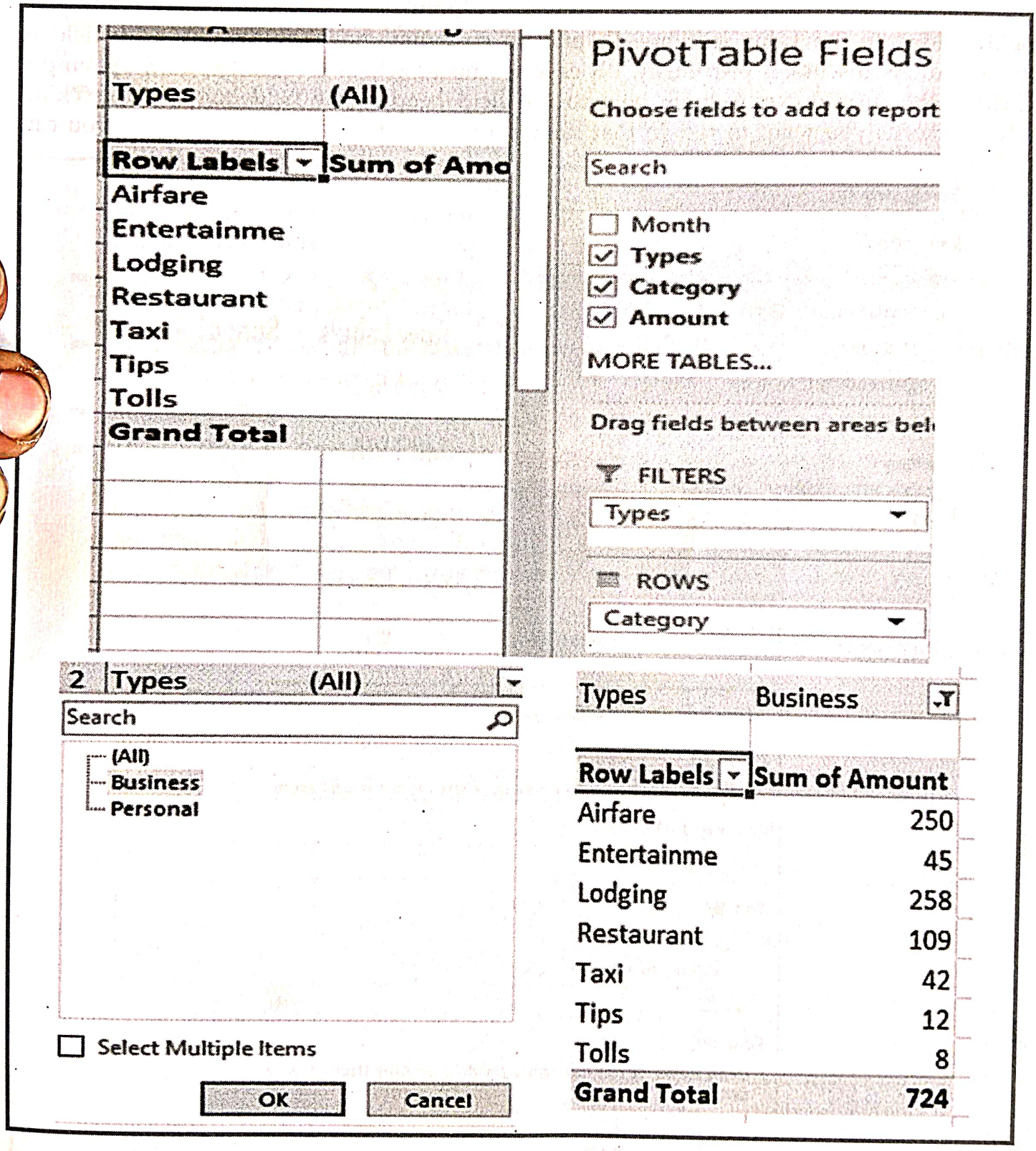
3. Click the item inside each slicer pane to display that data in the pivot table. (Hold down the Ctrl key and click to choose multiple items.)



10.23 FILTERING DATA IN PIVOT TABLE

To simplify your pivot tables, you may want to restrict them so that they show only a portion of the total data. This process is known as filtering. Report filtering lets you filter your data so that your pivot table uses only the rows that really interest you. For example, you may want to narrow your sales summary to a specific country. To do so, you need

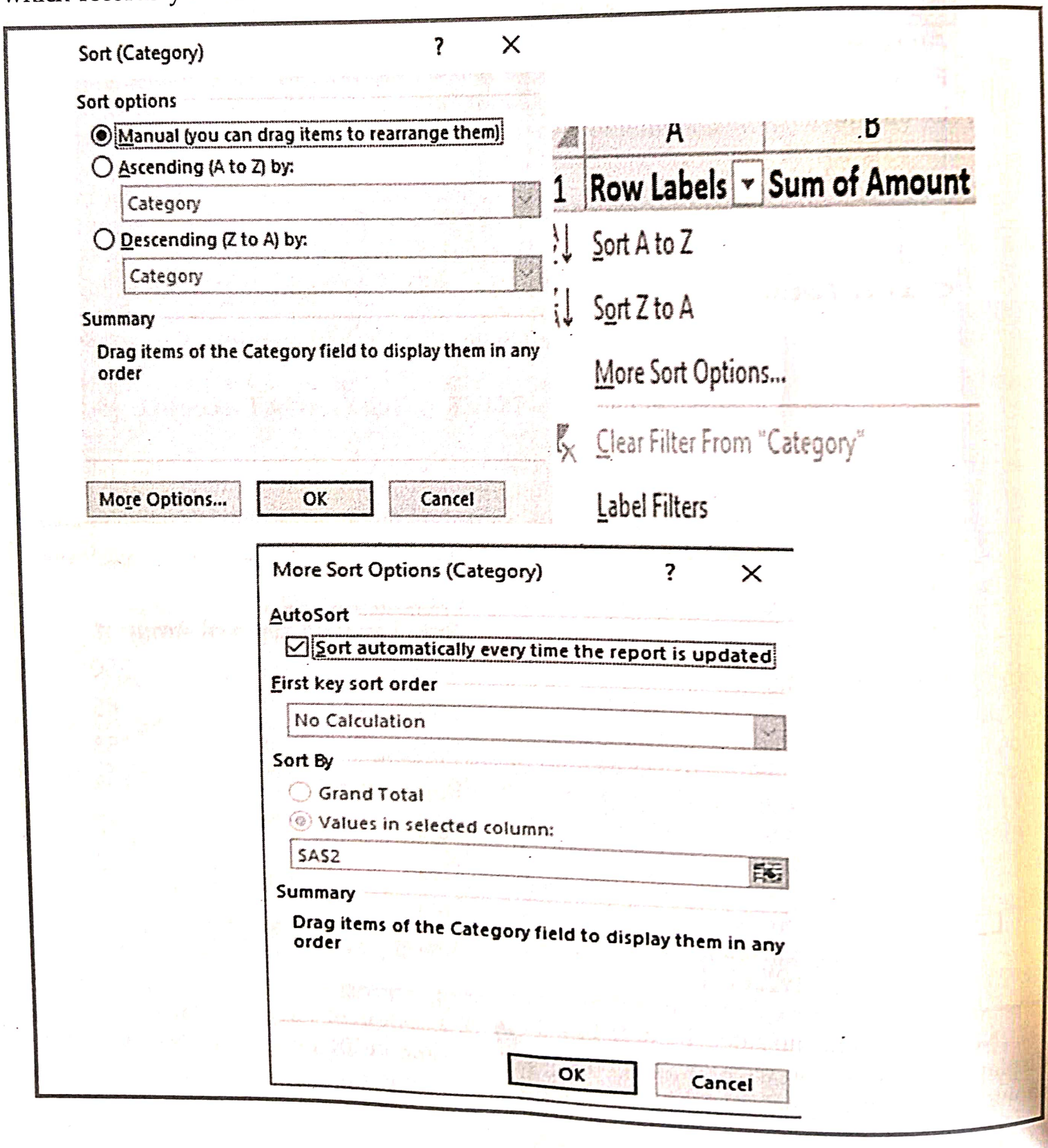
to drag the appropriate field (in this case, Ship Country), to the Filters box in the PivotTable Fields pane. You'll see it appear on your worksheet as shown in Figure.



Now Type field can filter the data based on the selection of the particular type. It will display the data of selected type and hide the other data in the pivot table. Second figure shows the filtered data.

10.24 SORTING DATA IN PIVOT TABLE

For each part of the PivotTable—row or column fields, the main filter field, even fields added to the value section—it's easy to customize how the data for each field is displayed. As discussed previously, it's easy to choose which records to show, by simply clicking the drop-down triangle next to the field header (the field name) and picking which records you want to see. In that same drop-down menu, shown in Figure, you can

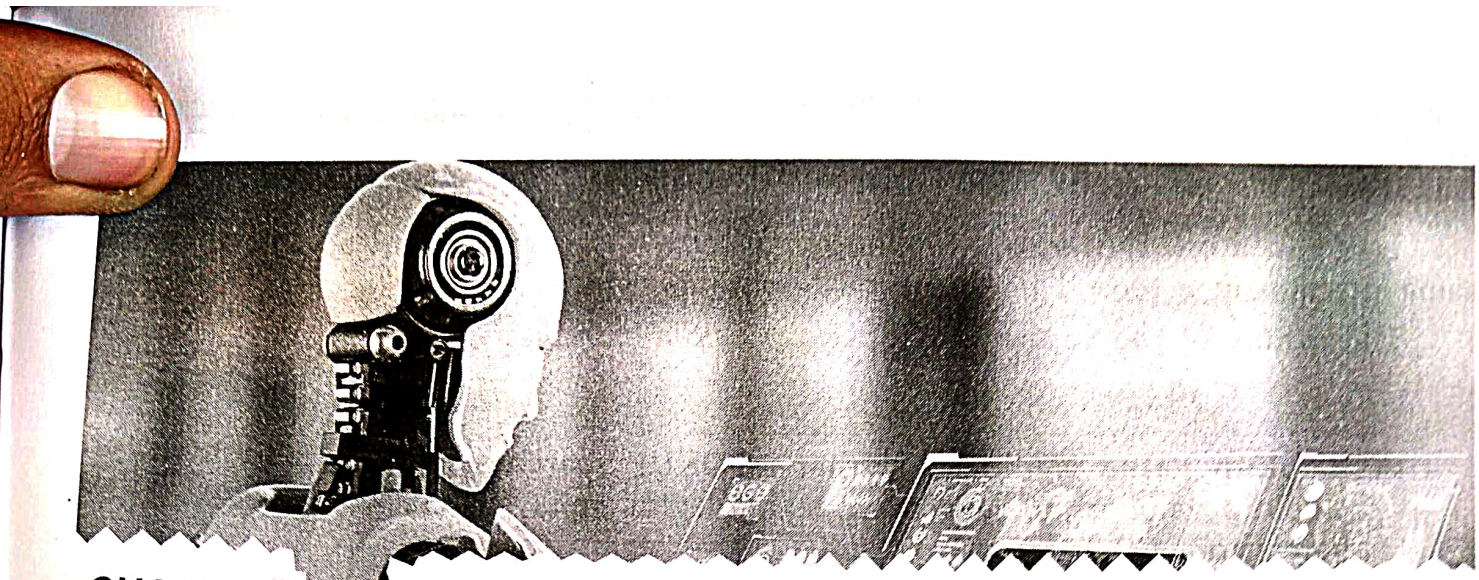


also choose to sort the records in A to Z(ascending) or Z to A (descending) order, or to choose More Sort Options.

The Sort dialog box (whose title bar includes the name of the field you were in at the time you opened the dialog box, as shown in Figure above, allows you to sort in the following ways:

- *Manual:* This first option is not on by default, but it can be handy if you don't want a traditional ascending or descending order sort. If you turn this option on, you are then free to drag records within the list (horizontal or vertical) of records for that field, creating a totally customized order. You might be ordering them by priority, preference, or some other quality that's not dictated by the first letter or a numerical value.
- *Ascending: (A to Z).* This option allows you to click the drop-down arrow to choose a field to sort in ascending order.
- *Descending: (Z to A).* This does the opposite with the field you select from the drop-down menu—it sorts it in descending order.
- *Summary:* This area simply describes the sort you've chosen.
- *More Options.* Click this button to display the dialog box shown in Figure, through which you can use AutoSort to automatically re-sort the report each time you update it with new data, or if you turn Auto Sort off, you can sort by calculated field values. You will rarely, if ever, need these options, so you'd only ever open this dialog box to access the Auto Sort option to turn it on or off.

☆☆☆☆☆



CHAPTER

11 SEARCHING USING HLOOK & VLOOK AND ANALYZING DATA

CHAPTER AT A GLANCE

11.1 Lookup Functions	11.3 Scenario Manager
11.2 Data Table (what-if Manually)	11.4 Goal Seek

11.1 LOOKUP FUNCTIONS

Getting the meaning of a word in the dictionary is always a two-step process: First you look up the word itself, and then you read its definition. Same with an encyclopedia: First look up the concept, and then read the article. This idea of looking something up to retrieve some related information is at the heart of many spreadsheet operations.

In many worksheet formulas, the value of one argument often depends on the value of another. Here are some examples:

- In a formula that calculates an invoice total, the customer's discount might depend on the number of units purchased.
- In a formula that charges interest on overdue accounts, the interest percentage might depend on the number of days each invoice is overdue.
- In a formula that calculates employee bonuses as a percentage of salary, the percentage might depend on how much the employee improved on the given budget. The usual way to handle these kinds of problems is to look up the appropriate value.

Note: For lookup function all columns of data must be sorted.

11.1.1. LOOKUP (VECTOR FORM) FUNCTION

Looks in a one-row or one-column range for a value and returns a value from the same position in a second one-row or one-column range.

The syntax of the vector form of the LOOKUP function is:

LOOKUP(lookup_value, lookup_vector, result_vector)

- **Lookup_value** is the value that LOOKUP searches for in the first vector. Lookup_value can be a number, text, a logical value, or a name or reference that refers to a value.
- **Lookup_vector** is a range that contains only one row or one column. The values in it can be text, numbers, or logical values. The values in lookup_vecto must be placed in ascending order:..., -2, -1, 0, 1, 2, ..., A-Z, FALSE, TRUE; otherwise, LOOKUP may not give the correct value. Uppercase text and lowercase text are considered equivalent.
- **Result_vector** is a range that contains only one row or column. It must be the same size as lookup_vector.

If LOOKUP cannot find lookup_value, it matches the largest value in lookup_vector that is less than or equal to lookup_value. This makes it possible to look up values where lookup_value falls in a range instead of matching a specific value. If lookup_value is smaller than the smallest value in lookup_vector, LOOKUP gives the #N/A! error value.

EXAMPLE: We continue with the example of Calculate share percent with of each month.

1			Jan	Feb	Mar
2		Profit	600	855	1073
3					
4	Partners	Share Percent	Jan Share	Feb Share	Mar Share
5	Dev	15%	90	128.25	160.95
6	Aman	18%	108	153.9	193.14
7	Arsh	20%	120	171	214.6
8	Guru	50%	300	427.5	536.5
9	Total	100%	600	855	1073

We use the lookup table to find the 'Jan Share' by entering the percentage of the parterres.

Here 'Share Percent' is lookup_vector from range to B5:B9. 'Jan Share' is result_vector from range to C5:C9.

=Lookup(18%, B5:B9, C5:C9) will return 108

=Lookup(50%, B5:B9, C5:C9) will return 300

=Lookup(25%, B5:B9, C5:C9) will return 120 as 25% is comes between 20% and 50% then it choose 20%.

D11 =LOOKUP(18%,B5:B9,C5:C9)

	A	B	C	D	E	F
1			Jan	Feb	Mar	
2		Profit	600	855	1073	
3						
4	Partners	Share Percent	Jan Share	Feb Share	Mar Share	
5	Dev	15%	90	128.25	160.95	
6	Aman	18%	108	153.9	193.14	
7	Arsh	20%	120	171	214.6	
8	Guru	50%	300	427.5	536.5	
9	Total	100%	600	855	1073	
10						
11				108		

D11 =LOOKUP(50%,B5:B9,C5:C9)

	A	B	C	D	E	F
1			Jan	Feb	Mar	
2		Profit	600	855	1073	
3						
4	Partners	Share Percent	Jan Share	Feb Share	Mar Share	
5	Dev	15%	90	128.25	160.95	
6	Aman	18%	108	153.9	193.14	
7	Arsh	20%	120	171	214.6	
8	Guru	50%	300	427.5	536.5	
9	Total	100%	600	855	1073	
10						
11				300		

D11 =LOOKUP(25%,B5:B9,C5:C9)

	A	B	C	D	E	F
1			Jan	Feb	Mar	
2		Profit	600	855	1073	
3						
4	Partners	Share Percent	Jan Share	Feb Share	Mar Share	
5	Dev	15%	90	128.25	160.95	
6	Aman	18%	108	153.9	193.14	
7	Arsh	20%	120	171	214.6	
8	Guru	50%	300	427.5	536.5	
9	Total	100%	600	855	1073	
10						
11				120		

11.1.2. HLOOKUP FUNCTION

The HLOOKUP() function is similar to VLOOKUP() except that it searches for the lookup value in the first row of a table. (The H in HLOOKUP() stands for horizontal.) If successful, this function then looks down the specified number of rows and returns the value it finds there. Here's the syntax for HLOOKUP():

HLOOKUP(lookup_value, table_array, row_index_num, [range_lookup])

- **lookup_value** argument is the value that you want to look up in the Lookup table.
- **table_array** is the cell range or name of the Lookup table that contains both the value to look up and the related value to return.
- **row_index_num** is the row number in table_array from which the matching value will be returned. A row_index_num of 1 returns the first row value in table_array, a row_index_num of 2 returns the second row value in table_array, and so on.
- The optional **range_lookup** argument in VLOOKUP is the logical TRUE or FALSE that specifies whether you want Excel to find an exact or approximate match for the lookup_value in the table_array. When you specify TRUE or omit the range_lookup argument in the VLOOKUP function, Excel finds an approximate match. When you specify FALSE as the range_lookup argument, Excel finds only exact matches.

EXAMPLE: Consider the following products data along with their sale price and cost.

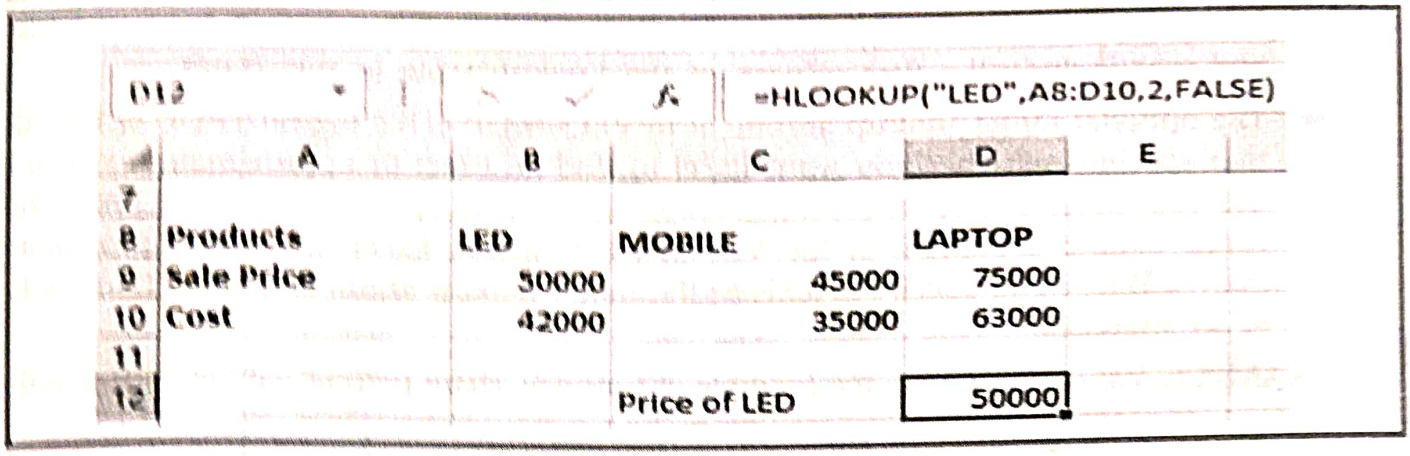
	A	B	C	D
7				
8	**Products**	**LED**	**MOBILE**	**LAPTOP**
9	**Sale Price**	50000	45000	75000
10	**Cost**	42000	35000	63000
11				

Using the HLOOKUP function let us get the following values from this table:

What is the price of a LED?

Ans. In this scenario lookup_value is "LED", table_array is from range A8:D10, row_index_num is 2 because we have to find price of the product and we want exact match of the product the FALSE will be used. The function will be

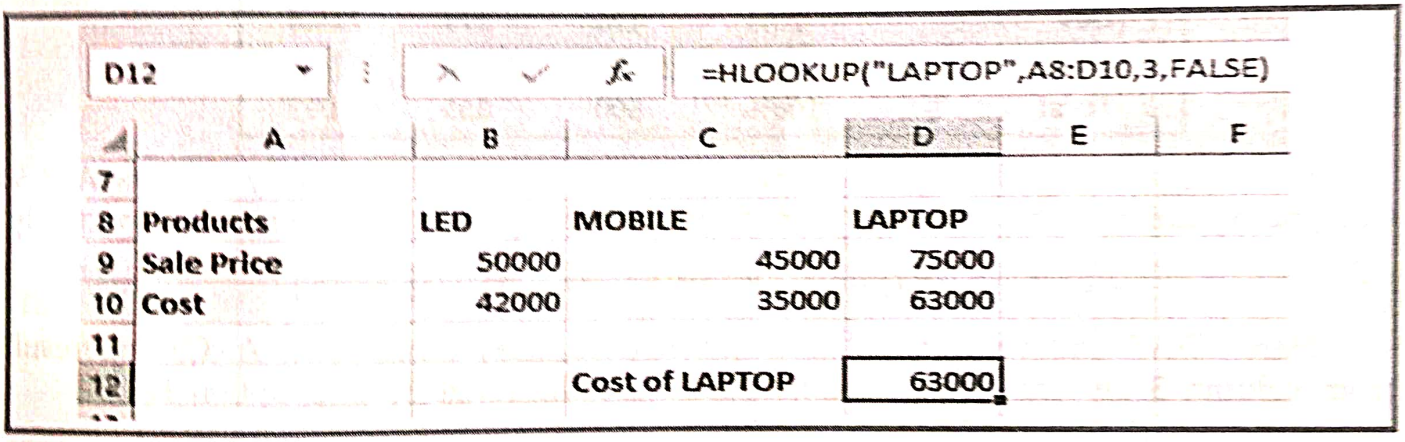
=HLOOKUP("LED", A8:D10, 2, FALSE)



What is the cost of a LAPTOP?

Ans. In this scenario lookup_value is "LAPTOP", table_array is from range A8:D10, row_index_num is 3 because we have to find the cost of the product and we want exact match of the product the FALSE will be used. The function will be

=HLOOKUP("LAPTOP", A8:D10, 3, FALSE)



11.1.3. VLOOKUP FUNCTION

The VLOOKUP function is used to perform a vertical look up in the left most column and return a value in the same row from a column that you specify. It then looks across a specified number of columns and returns whatever value it finds there.

The VLOOKUP function uses the following syntax:

VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])

- **lookup_value** argument is the value that you want to look up in the Lookup table.
- **table_array** is the cell range or name of the Lookup table that contains both the value to look up and the related value to return.
- **col_index_num** argument designates the column of the lookup table containing the values that are returned by the VLOOKUP function based on matching the value of the lookup_value argument against those in the table_ array argument. You determine the col_index_num argument counting how many columns this

column is over to the right from the first column of the vertical Lookup table, and you include the first column of the Lookup table in this count.

- The optional **range_lookup** argument in VLOOKUP is the logical TRUE or FALSE that specifies whether you want Excel to find an exact or approximate match for the lookup_value in the table_array. When you specify TRUE or omit the range_lookup argument in the VLOOKUP function, Excel finds an approximate match. When you specify FALSE as the range_lookup argument, Excel finds only exact matches.

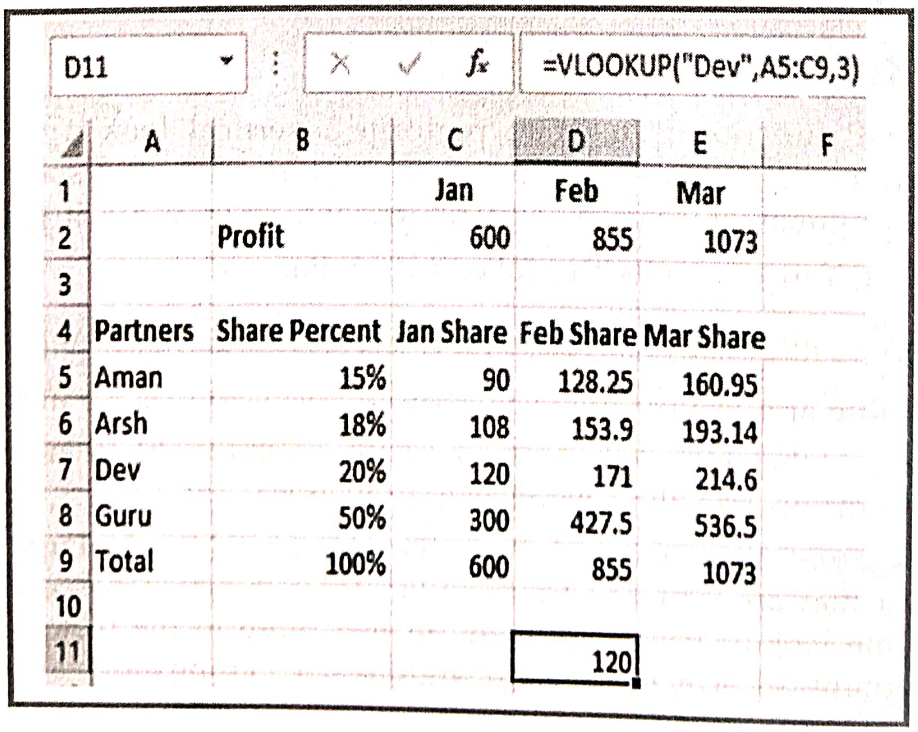
EXAMPLE: We continue with the example of Calculate share percent with of each month.

	A	B	C	D	E
1			Jan	Feb	Mar
2		Profit	600	855	1073
3					
4	Partners	Share Percent	Jan Share	Feb Share	Mar Share
5	Aman	15%	90	128.25	160.95
6	Arsh	18%	108	153.9	193.14
7	Dev	20%	120	171	214.6
8	Guru	50%	300	427.5	536.5
9	Total	100%	600	855	1073

Here we are considering 3 columns A, B and C only. Table_array is from A5:C9. CoL_index_num is 2 or 3 based on your choice i.e 'Share percent' or 'Jan share' respectively.

= =VLOOKUP("Dev", A5:C9, 3) will return 120.

Here "Dev" will be searched from given table_array range from A5:C9 and result from column 3 i.e. 'Jan share'. Second figure use the range_lookup option FALSE



D11 =VLOOKUP("Dev",A5:C9,3,FALSE)

	A	B	C	D	E	F
1			Jan	Feb	Mar	
2		Profit	600	855	1073	
3						
4	Partners	Share Percent	Jan Share	Feb Share	Mar Share	
5	Aman	15%	90	128.25	160.95	
6	Arsh	18%	108	153.9	193.14	
7	Dev	20%	120	171	214.6	
8	Guru	50%	300	427.5	536.5	
9	Total	100%	600	855	1073	
10						
11				120		

If exact match is not found and if the range_lookup is TRUE

D11 =VLOOKUP("De",A5:C9,3,FALSE)

	A	B	C	D	E	F	G
1			Jan	Feb	Mar		
2		Profit	600	855	1073		
3							
4	Partners	Share Percent	Jan Share	Feb Share	Mar Share		
5	Aman	15%	90	128.25	160.95		
6	Arsh	18%	108	153.9	193.14		
7	Dev	20%	120	171	214.6		
8	Guru	50%	300	427.5	536.5		
9	Total	100%	600	855	1073		
10							
11				#N/A			

D11 =VLOOKUP("Gu",A5:C9,3,TRUE)

	A	B	C	D	E	F
1			Jan	Feb	Mar	
2		Profit	600	855	1073	
3						
4	Partners	Share Percent	Jan Share	Feb Share	Mar Share	
5	Aman	15%	90	128.25	160.95	
6	Arsh	18%	108	153.9	193.14	
7	Dev	20%	120	171	214.6	
8	Guru	50%	300	427.5	536.5	
9	Total	100%	600	855	1073	
10						
11				120		

11.2 DATA TABLE (WHAT-IF MANUALLY)

Changing an input value in an Excel spreadsheet can affect the result of a formula. Simply insert the new value in the cell that feeds into the formula. When you alter the input value, Excel will recalculate the calculation and display the updated results. This method is useful for undertaking what-if analysis and comparing results across multiple input values in a single worksheet.

This type of what-if analysis can be performed using Excel's Data Table command. To create a data table, add a sequence of input values in the worksheet. Excel will use each value in the specified formula. When Excel finishes computing the data table, it displays the results for each change in input values in a specific range of the worksheet. You can save the data table as a worksheet to record the results of several input values.

When constructing data tables, you can create a one-variable or a two-variable data table. In a one-variable data table, Excel substitutes several values for a single input value in a formula. Excel uses a formula to replace two input values in a two-variable data table.

11.2.1 INDEPENDENT AND DEPENDENT VARIABLES

Independent variables are the model's inputs. They are not calculated in the model; the user must provide them. The term "independent" refers to the ability for users to make independent changes. The independent variable fields in worksheets solely accept input numbers or text, without any formulas.

In a model, the value of a dependent variable is determined by one or more independent variables, but not necessarily by all of them. When using an Excel feature that assumes a dependent variable (x) depends on a specific independent variable (y).

Since its value is dependent on the values of other independent and dependent variables, any value computed within the model is a dependent variable. While some dependent variables may be intermediate values used to compute other dependent variables, others are the model's final outputs. In the worksheets, any cell that has a formula in it represents a dependent variable.

11.2.2 ONE-VARIABLE DATA TABLE

Setting up the master formula in your worksheet and then entering a range of values that you wish to replace a single input value in that formula will allow you to create a one-variable data table.

In the present time, we are used to purchase the things online from e-commerce sites. Before purchasing the item, we are keen interested to know about EMI for the product with their different payment terms. Following example shows how one-variable table is constructed to display data of this type of query. This example discusses the purchasing of mobile of with payment terms and given interest rate then it calculates the EMI of the given data.

This example uses the PMT method to calculate the EMI where mobile price, total payment terms and annually interest rate is provided from cell B1:B3. Then write the

PMT method in cell B4 as shown in the picture to calculate the EMI amount as shown in the picture below.

B4 | =PMT(B3/12,B2,-B1)

	A	B	C	D	E
1	Mobile Price	₹ 15,000.00			
2	Loan Term	12			
3	Loan Interest	5%			
4	EMI	₹ 1,284.11			

To build the one-variable data table we can query one of the followings

- What will be the EMI if we change mobile price?
- What will be the EMI if we change the loan payment terms?
- What will be the EMI if we change the annual loan interest rate?

We will build the one-variable data table one by one to answer these queries separately.

What will be the EMI if we change mobile price?

1. Construct a table which has different mobile price but loan term and loan interest remains same as shown below.

B5 | =PMT(B4/12,B3,-B2)

	A	B	C	D	E
1	Calculate EMI with Different Mobile Price				
2	Mobile Price	₹ 15,000.00	₹ 25,000.00	₹ 30,000.00	₹ 45,000.00
3	Loan Term	12	12	12	12
4	Loan Interest	5%	5%	5%	5%
5	EMI	₹ 1,284.11			

2. Now calculate the EMI with PMT method in cell B5 and the drag to all cells then in the right to build one-variable data table.

	A	B	C	D	E
1	Calculate EMI with Different Mobile Price				
2	Mobile Price	₹ 15,000.00	₹ 25,000.00	₹ 30,000.00	₹ 45,000.00
3	Loan Term	12	12	12	12
4	Loan Interest	5%	5%	5%	5%
5	EMI	₹ 1,284.11	₹ 2,140.19	₹ 2,568.22	₹ 3,852.34

What will be the EMI if we change the loan payment terms?

1. Construct a table which has different loan terms but mobile price, loan interest remains same as shown below.

B5 : =PMT(B4/12,B3,-B2)

	A	B	C	D	E
1	Calculate EMI with Different Loan Term				
2	Mobile Price	₹ 15,000.00	₹ 15,000.00	₹ 15,000.00	₹ 15,000.00
3	Loan Term	12	18	24	36
4	Loan Interest	5%	5%	5%	5%
5	EMI	₹ 1,284.11			

2. Now calculate the EMI with PMT method in cell B5 and the drag to all cells then in the right to build one-variable data table.

	A	B	C	D	E
1	Calculate EMI with Different Loan Term				
2	Mobile Price	₹ 15,000.00	₹ 15,000.00	₹ 15,000.00	₹ 15,000.00
3	Loan Term	12	18	24	36
4	Loan Interest	5%	5%	5%	5%
5	EMI	₹ 1,284.11	₹ 866.71	₹ 658.07	₹ 449.56

What will be the EMI if we change the annual loan interest rate?

1. Construct a table which has different loan interest but loan term and mobile price remains same as shown below.

	A	B	C	D	E
1	Calculate EMI with Different Loan Interest				
2	Mobile Price	₹ 15,000.00	₹ 15,000.00	₹ 15,000.00	₹ 15,000.00
3	Loan Term	12	12	12	12
4	Loan Interest	5%	7%	8%	10.5%
5	EMI	₹ 1,284.11			

2. Now calculate the EMI with PMT method in cell B5 and the drag to all cells then in the right to build one-variable data table.

Calculate EMI with Different Loan Interest				
Mobile Price	₹ 15,000.00	₹ 15,000.00	₹ 15,000.00	₹ 15,000.00
Loan Term	12	12	12	12
Loan Interest	5%	7%	8%	10.5%
EMI	₹ 1,284.11	₹ 1,294.45	₹ 1,304.83	₹ 1,322.23

11.2.3 TWO-VARIABLE DATA TABLE

A two-variable data table is created when you have a master formula in a worksheet and you wish to observe the impact of altering two of its input values. A single-row range in the table's first row and a single-column range in the first column are the two ranges of input values that are entered to be replaced in the master formula when creating a two-variable data table. A copy of the master formula is inserted into the cell where this row and column of input values overlap when creating a two-variable data table.

Continue with the previous example if we have queries having two different values at a time and we want to see the result of both different values then build two-variable data table.

For example, what is the EMI if we have different mobile price and Loan payment while loan interest remains same then calculate the EMI in two-variable data table as follows.

1. Create a table with different mobile price in different rows and loan payment terms in different columns range from A2:E6 and loan interest is in the F2 cell.

	A	B	C	D	E	F
1	Calculate EMI with Different Mobile Price and Loan Terms					Loan Interest
2	Mobile Price/Loan Terms	12	18	24	36	5%
3	₹ 15,000.00					
4	₹ 20,000.00					
5	₹ 35,000.00					

2. Click on the B3 cell and write the PMT formula as shown in the figure. Please note that we have used absolute reference for the Loan Interest as it must

B3 : =PMT(F2/12,B$2,-$A3)

	A	B	C	D	E
1	Calculate EMI with Different Mobile Price and Loan Terms				
2	Mobile Price/Loan Terms	12	18	24	36
3	₹ 15,000.00	₹ 1,284.11			
4	₹ 20,000.00				
5	₹ 35,000.00				

remain same when formula will be used in other cells. Mixed cell reference is used for Mobile price and Loan Terms. For mobile price Column must be same but row should be change. But in Loan Terms row must remain same with changing column. Examine the formula carefully in the figure given below.

3. Now drag the formula in rows and column to fill the entire table.

	A	B	C	D	E
1	Calculate EMI with Different Mobile Price and Loan Terms				
2	Mobile Price/Loan Terms	12	18	24	36
3	₹ 15,000.00	₹ 1,284.11	₹ 866.71	₹ 658.07	₹ 449.56
4	₹ 20,000.00	₹ 1,712.15	₹ 1,155.61	₹ 877.43	₹ 599.42
5	₹ 35,000.00	₹ 2,996.26	₹ 2,022.32	₹ 1,535.50	₹ 1,048.98

Try to build two-variable data table using different combinations (mobile price and loan interest or loan terms and loan interest and so on) to understand the concept in better way.

11.2.4 LIMITATION OF DATA TABLE

Data tables are valuable, however there are several limitations:

- You can only adjust one or two input cells at a time.
- Creating a data table is not intuitive.
- A two-input table displays the results of a single formula cell. Additional tables can be created for multiple formulas.
- Often, you just need a few possibilities of two input cells, rather than a full table.

11.3 SCENARIO MANAGER

The Data Table command is ideal for models with two variables that can vary. However, some models have multiple variables that can vary. You can use the Scenario Manager to build numerous scenarios with up to 32 variables each.

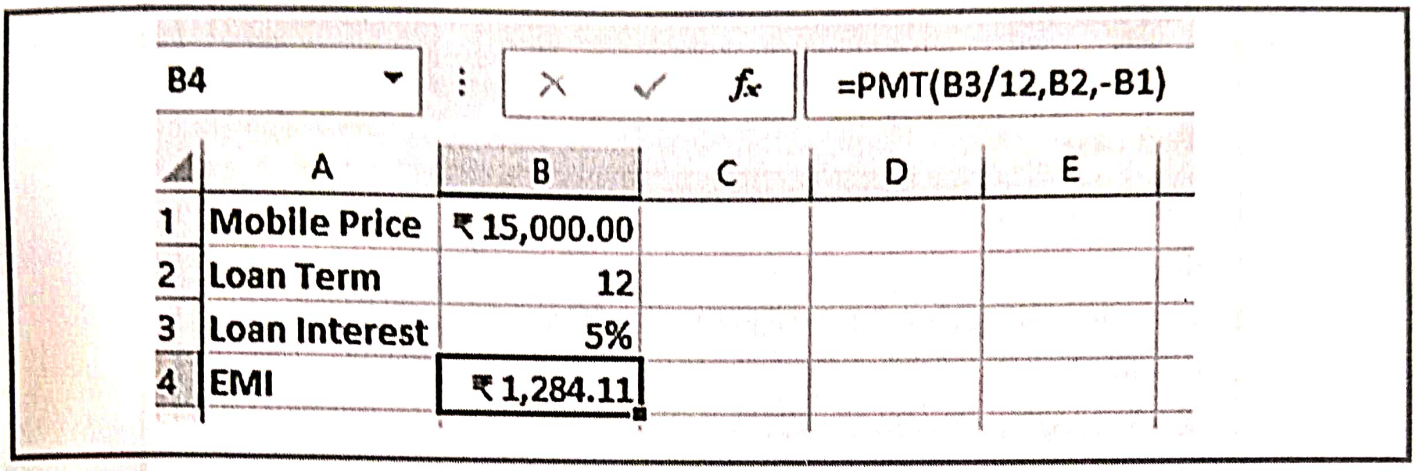
To handle up to 32 variable changes, utilize named ranges for all input variables before defining the first scenario. The Scenario Manager generates a summary report. Using named ranges for all input cells improves the report's clarity.

A scenario is a group of input values in a worksheet with a name (e.g., Best Case, Worst Case, Most Likely Case). To reuse input data and examine outcomes on the worksheet, pick the desired scenario and Excel will apply the stored input values to the appropriate cells.

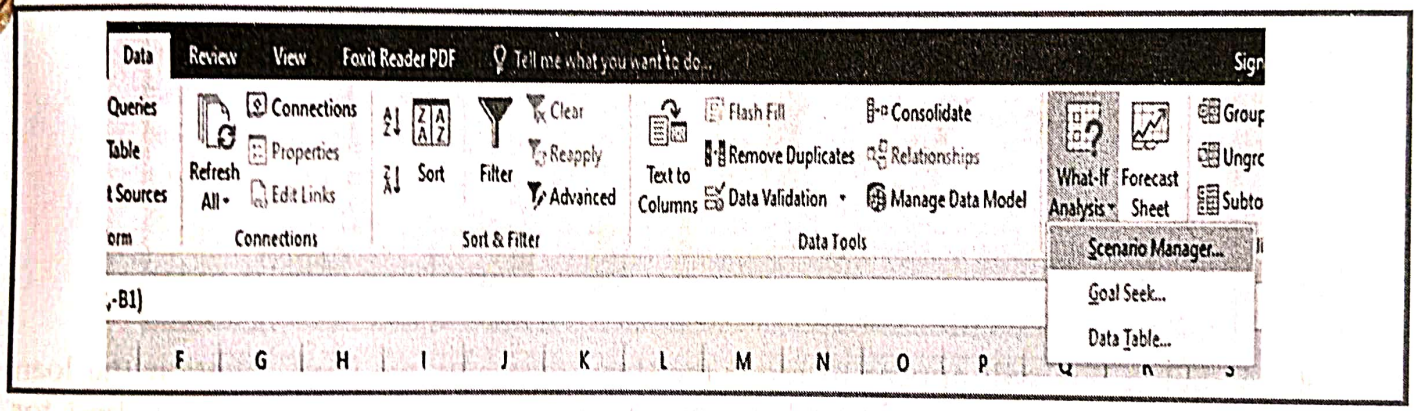
To construct a scenario for your worksheet, create a spreadsheet with dynamic cells. To easily identify changing cells in your scenarios. To construct this table, manually change the input data in the model's cells and copy the values of the dependent variables into a table, similar to creating a data table. The Scenario Manager automates the process, saving you valuable time.

11.3.1 DEFINING SCENARIO MANAGER FOR ONE VARIABLE

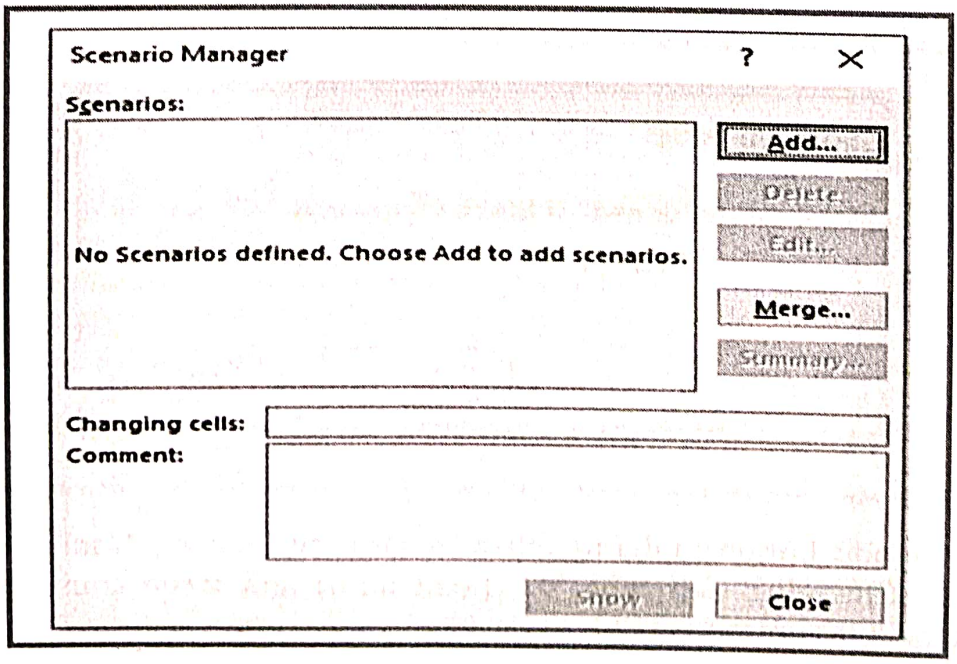
Consider the loan EMI example discussed in the previous section. For simplicity we define the scenario manager with single variable using different case. In this case we use different loan terms as a single variable.



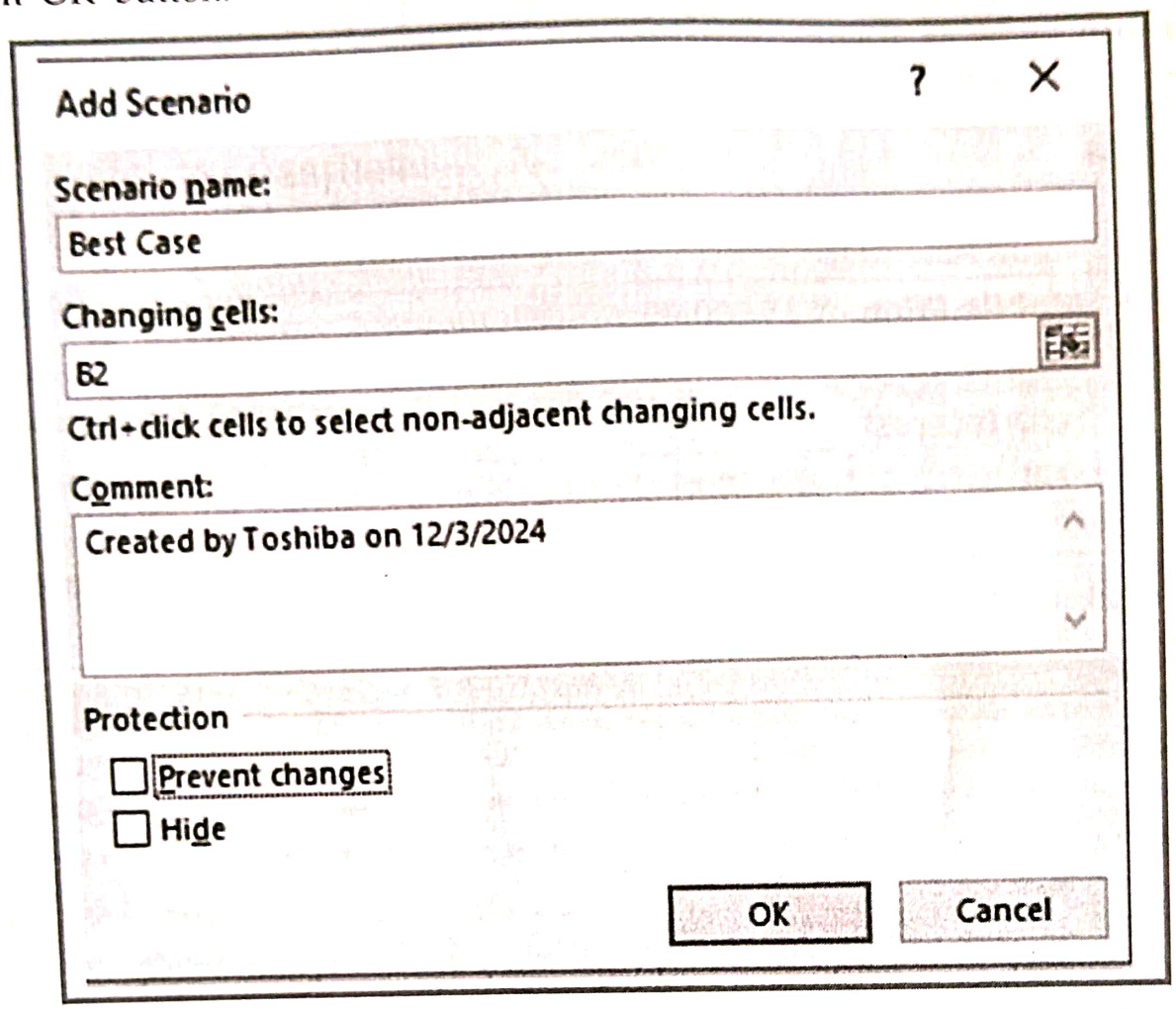
1. Click on Data Tab -> What-if Analysis-> Scenario Manager



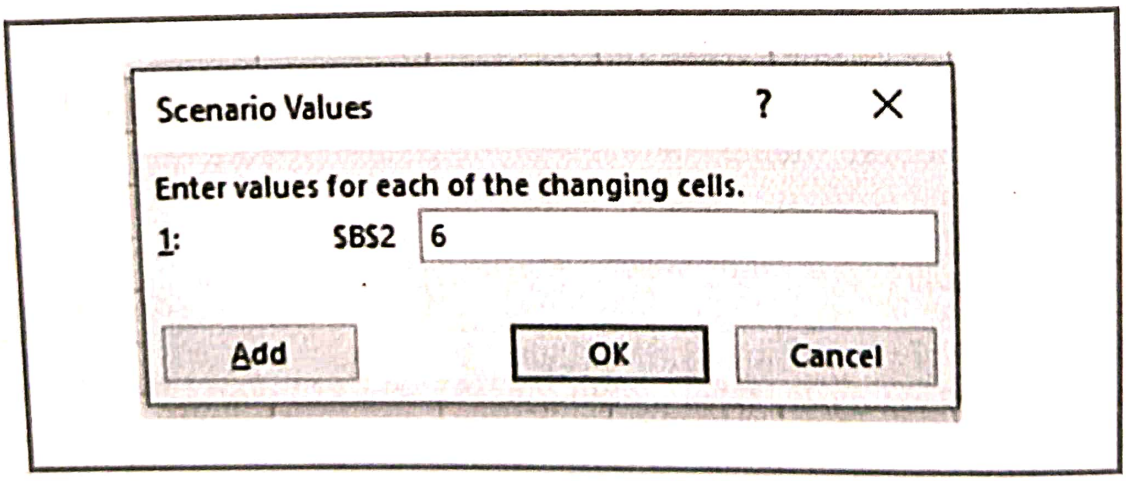
2. Scenario Manager Window is appeared.



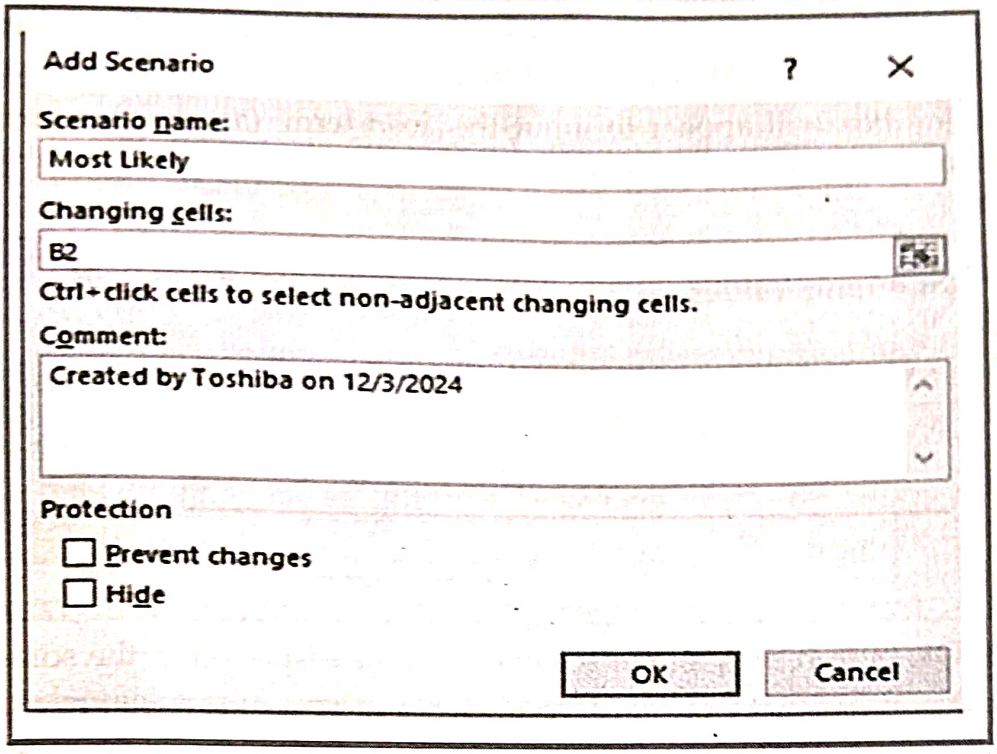
3. Click on Add button to create new scenario. In Scenario Name field type 'Best Case', Changing Cell type 'B2' (loan term) and write comment if required. Click on OK button.



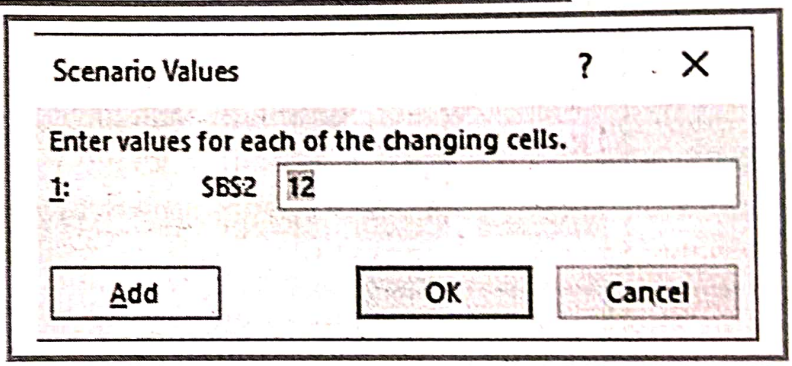
4. A dialog box will appear to input the 'loan term' (as we have mention the loan term as changing cell). Enter loan term 6 as we consider small terms is best for EMI.



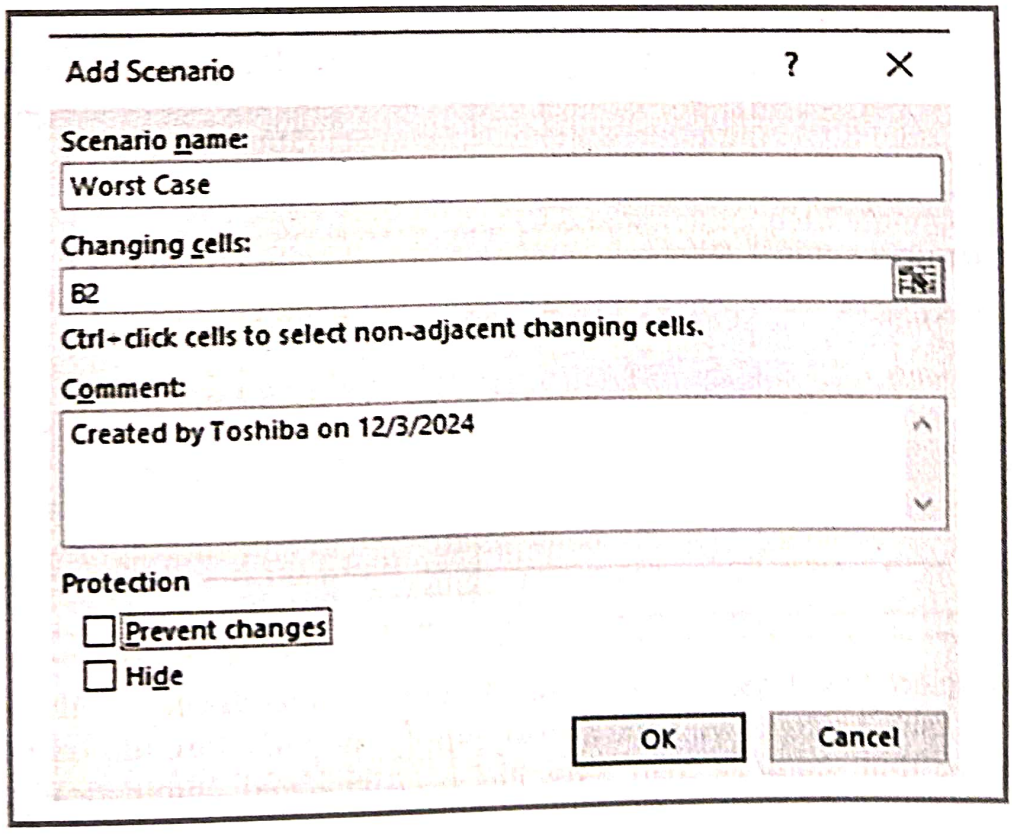
5. Click on add button to define other scenario. In Scenario Name field type 'Most Likely', Changing Cell type 'B2' (loan term) and write comment if required. Click on OK button.



6. A dialog box will appear to input the 'loan term' (as we have mention the loan term as changing cell). Enter loan term 12 as we consider this term is Mostly liked for EMI.

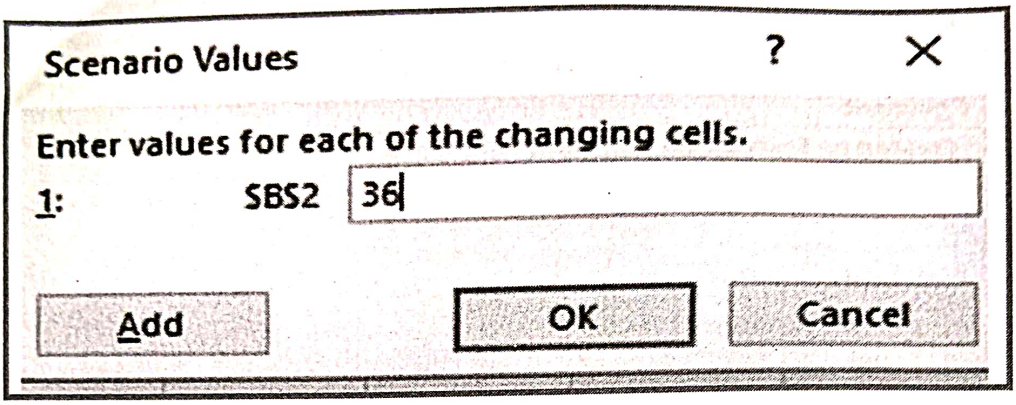


7. Click on add button to define other scenario. In Scenario

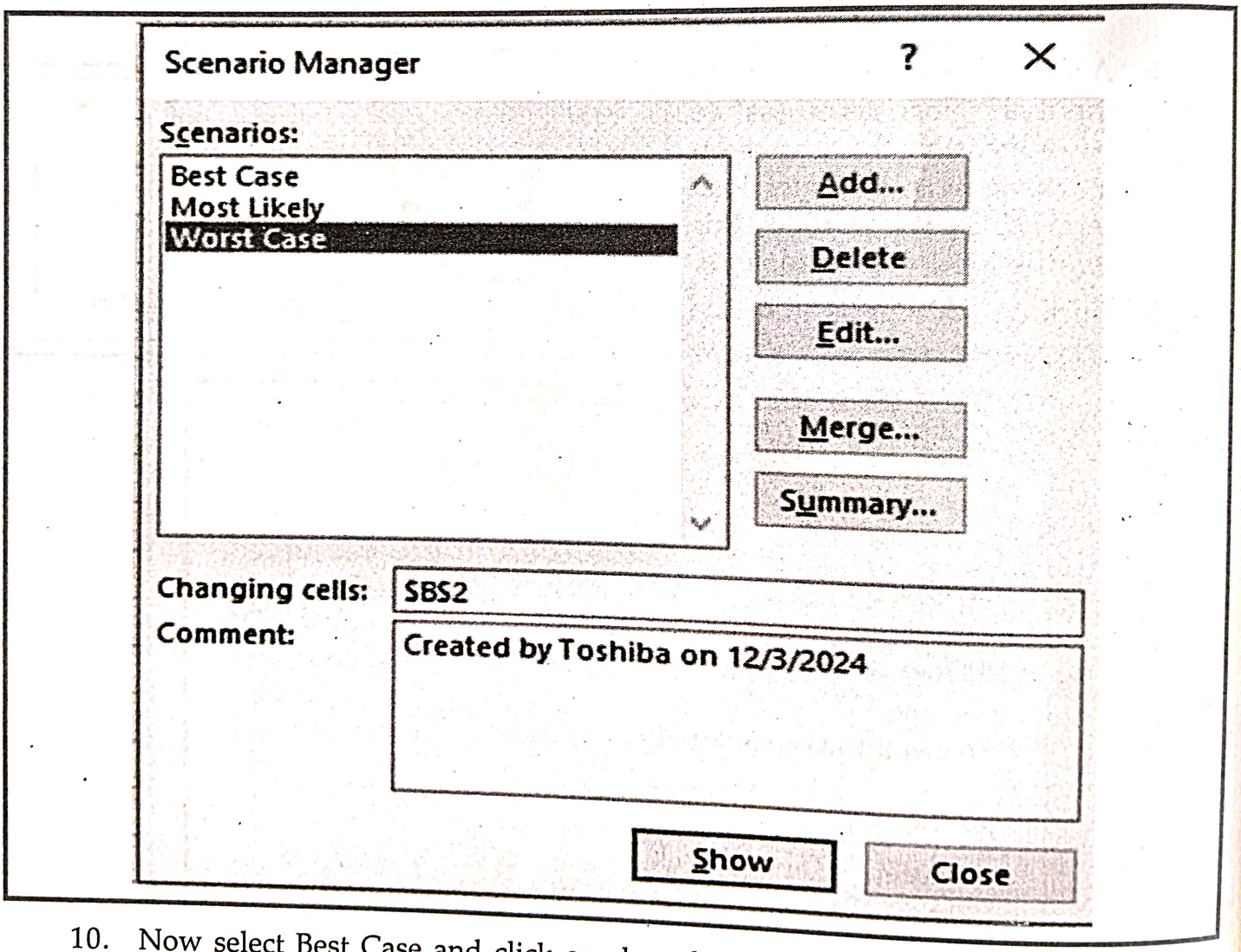


Name field type 'Worst Case', Changing Cell type 'B2' (loan term) and write comment if required. Click on OK button.

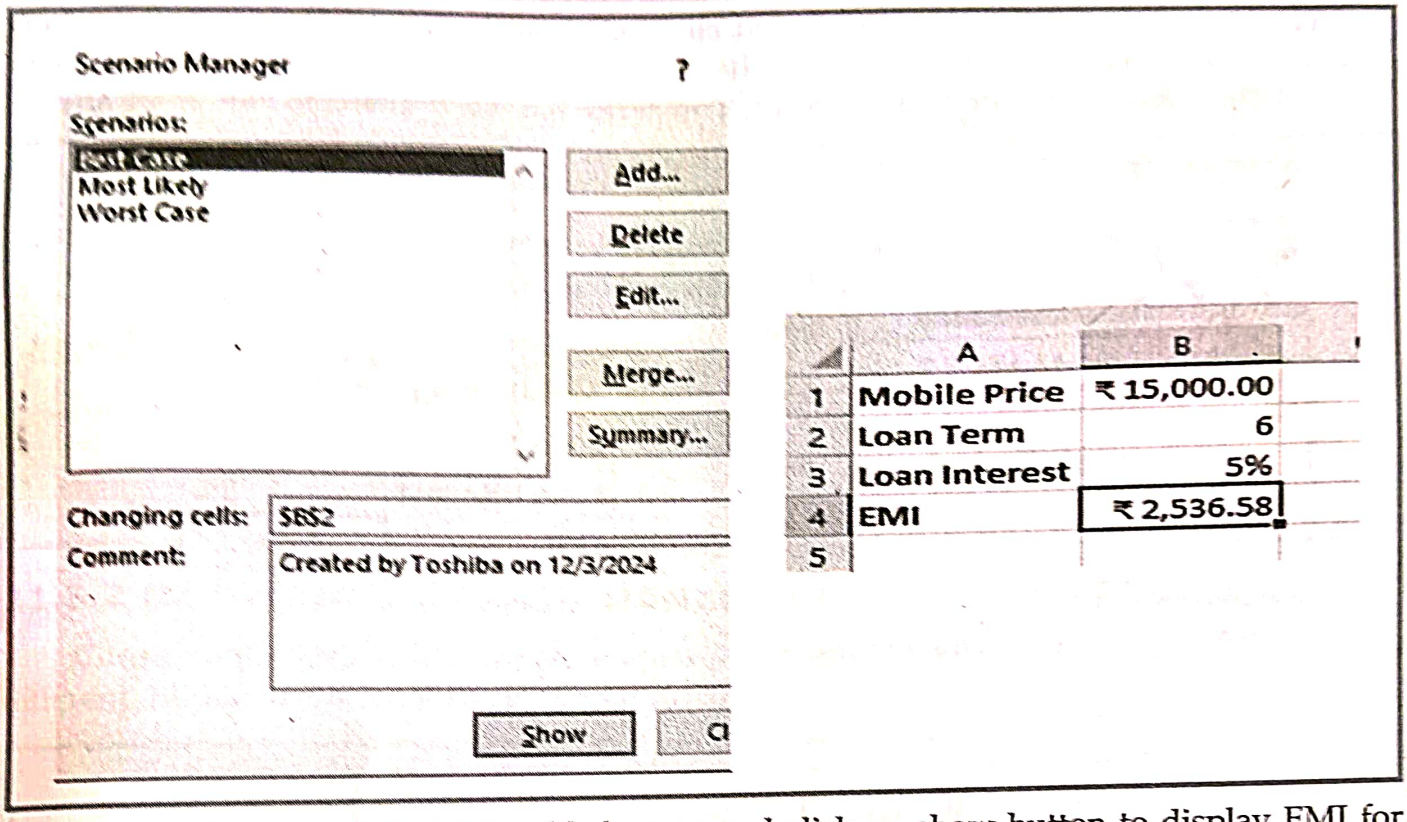
8. A dialog box will appear to input the 'loan term' (as we have mention the loan term as changing cell). Enter loan term 36 as we consider this terms as Worst EMI.



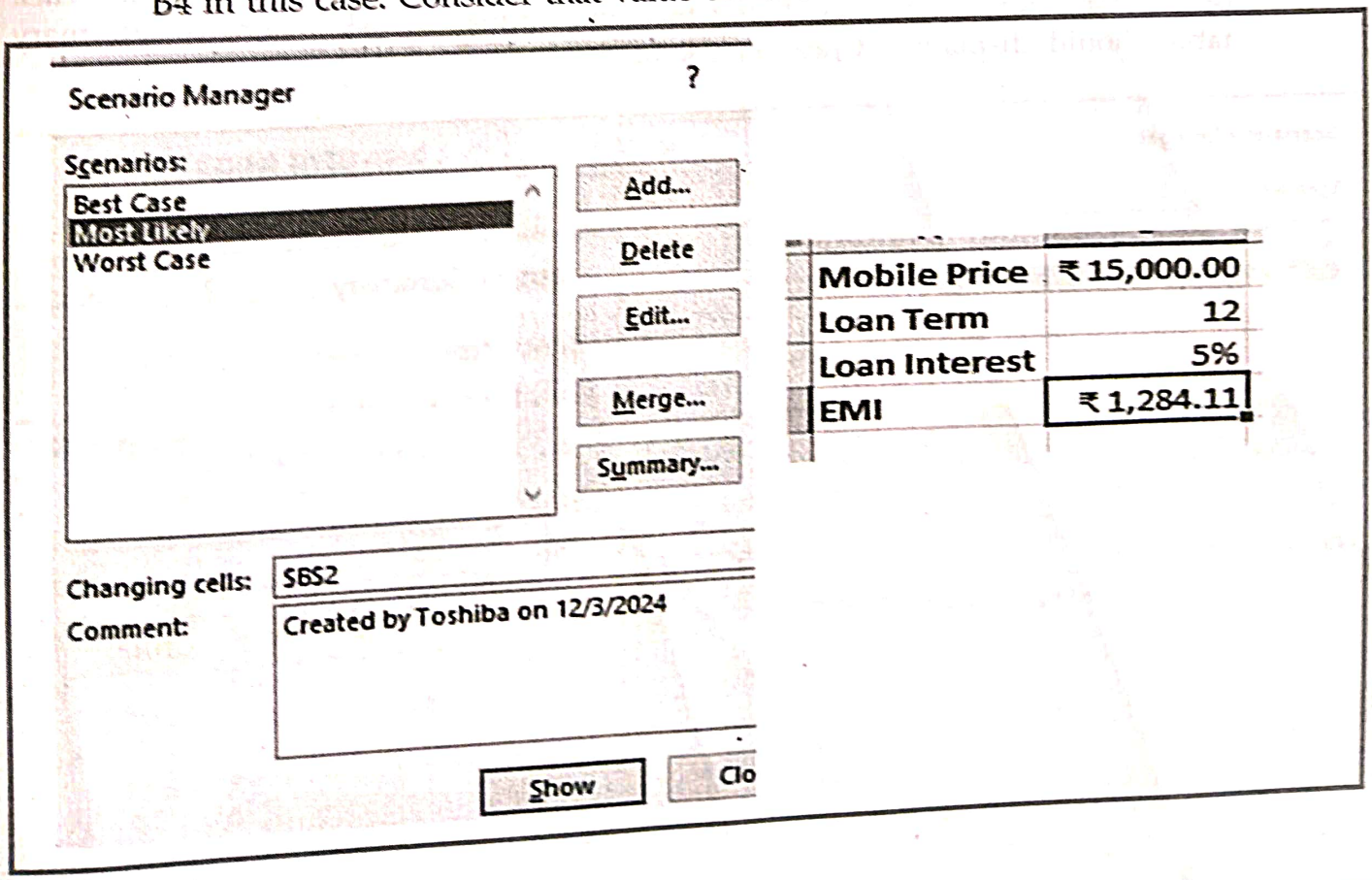
9. Click OK button then all the scenarios will be displayed in the scenario manager.



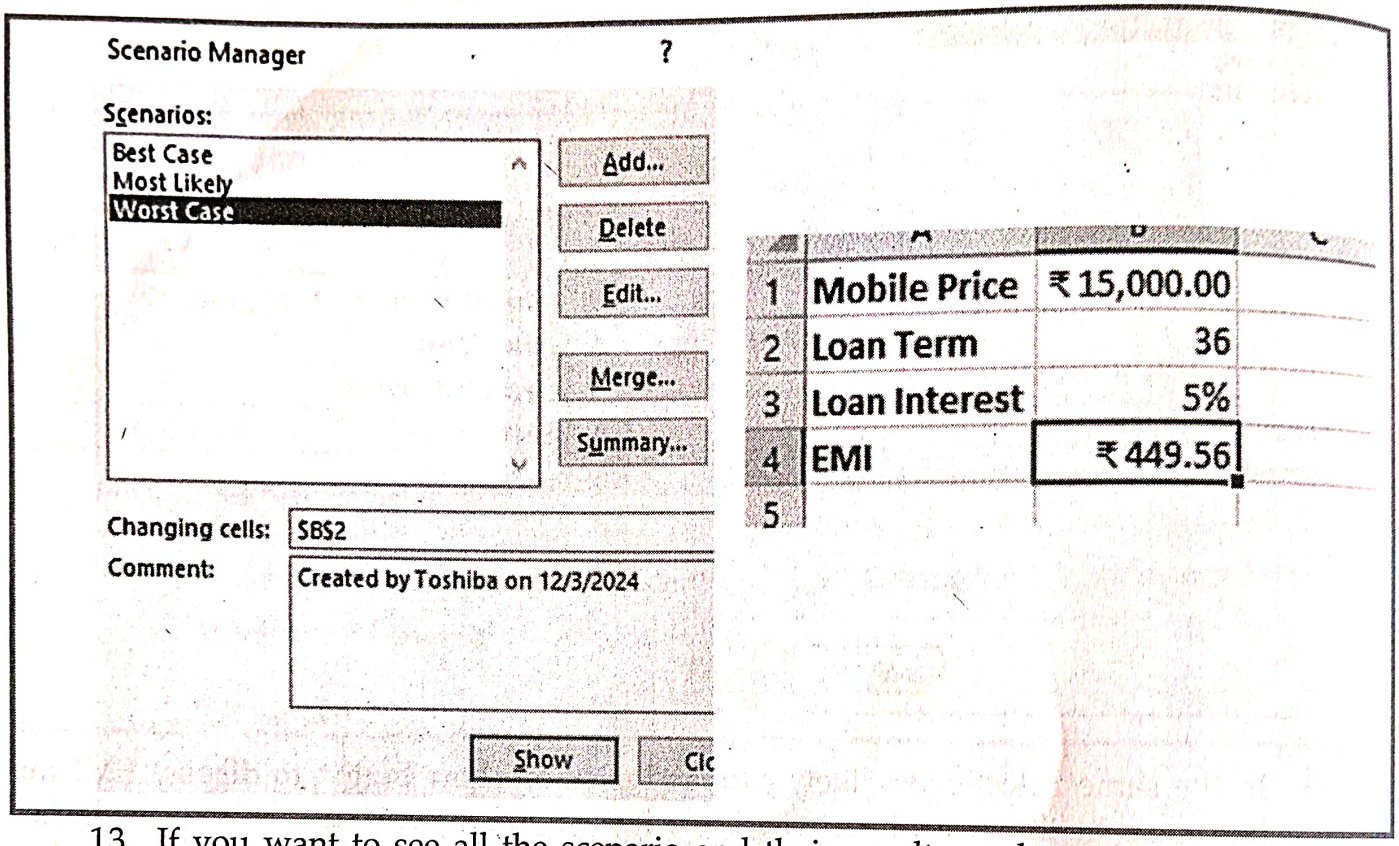
10. Now select Best Case and click on show button to display EMI for best case in the worksheet where we have already apply the PMT formula i.e. B4 in this case. Consider that value of loan term has been changed automatically.



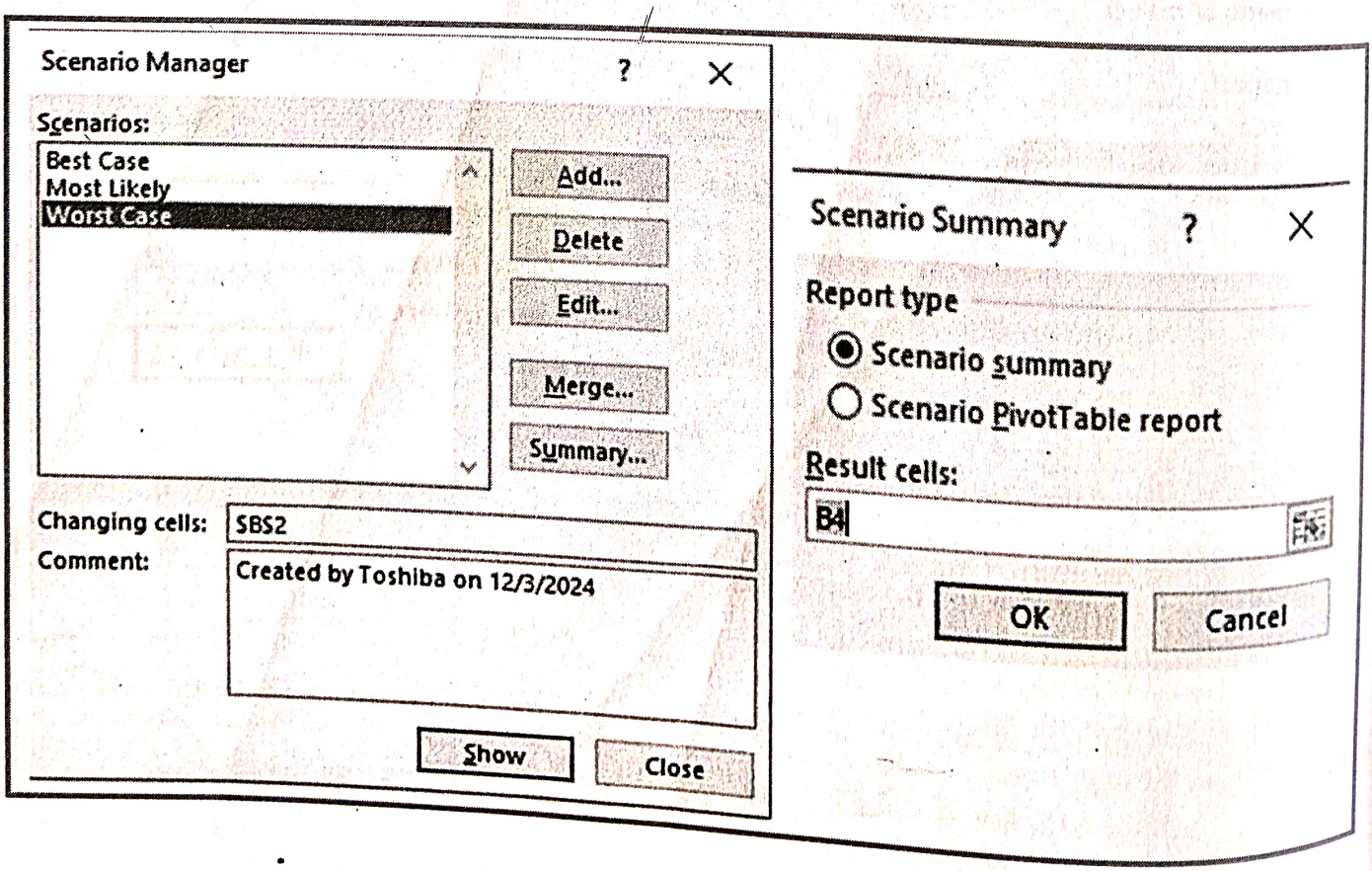
11. In the same Select Most likely case and click on show button to display EMI for best case in the worksheet where we have already apply the PMT formula i.e. B4 in this case. Consider that value of loan term has been changed automatically.



12. In the same Select Worst case and click on show button to display EMI for best case in the worksheet where we have already apply the PMT formula i.e. B4 in this case. Consider that value of loan term has been changed automatically.



13. If you want to see all the scenario and their results at the same time, then click on the summary button. A dialog box will appear to ask where to summary table should displayed. Click on Ok button.



14. Following is the summary table of all the scenarios.

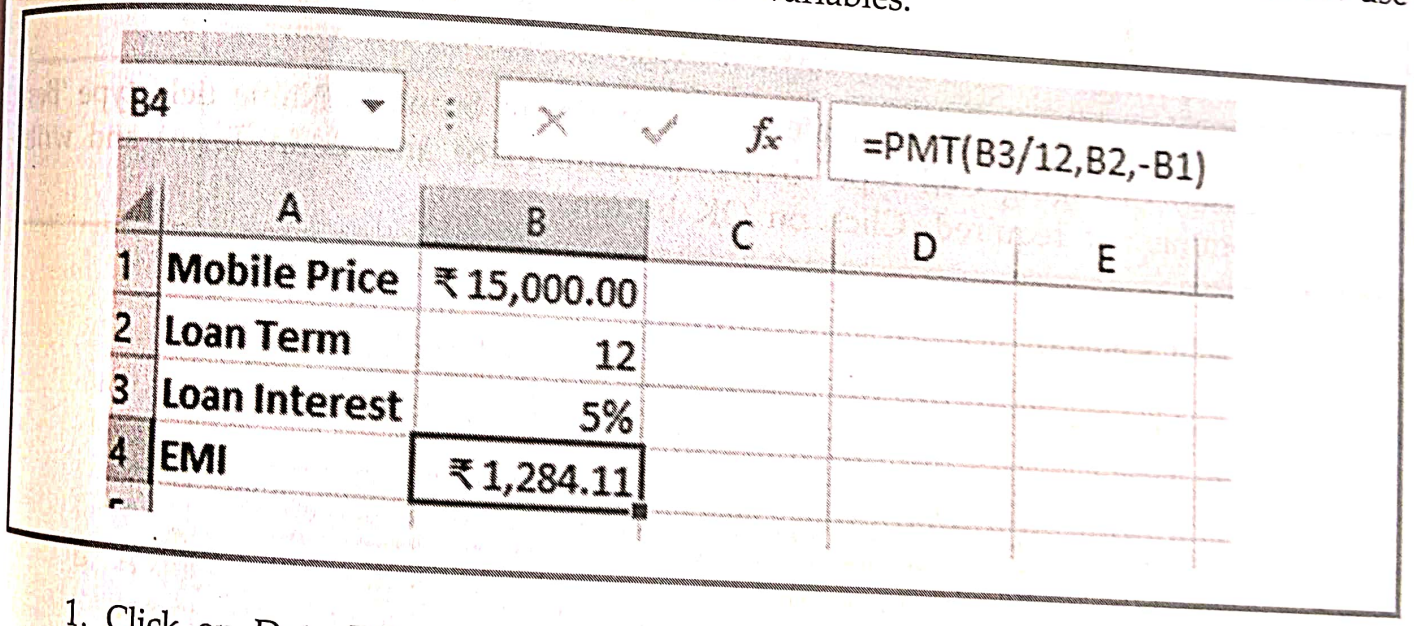
Scenario Summary

	Current Values:	Best Case	Most Likely	Worst Case
Changing Cells:				
B2	36	6	12	36
Result Cells:				
B4	₹449.56	₹2,536.58	₹1,284.11	₹449.56

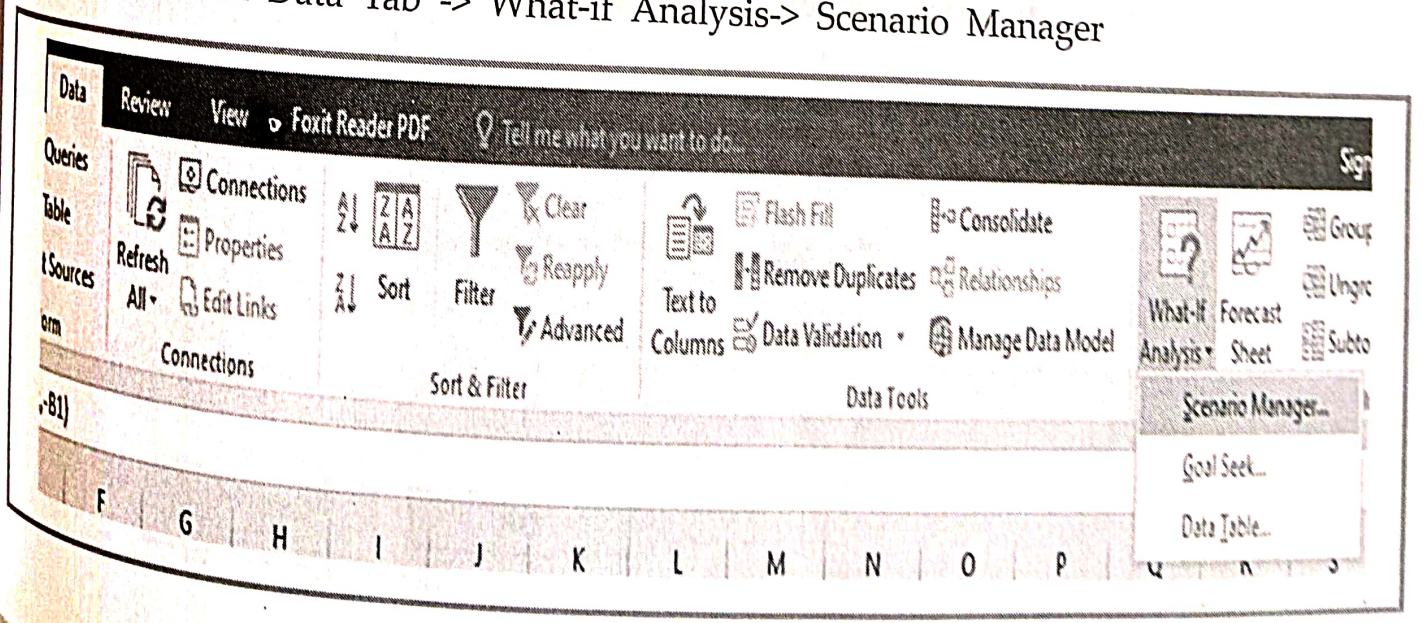
Notes: Current Values column represents values of changing cells at time Scenario Summary Report was created. Changing cells for each scenario are highlighted in gray.

11.3.2 DEFINING SCENARIO MANAGER FOR TWO VARIABLES

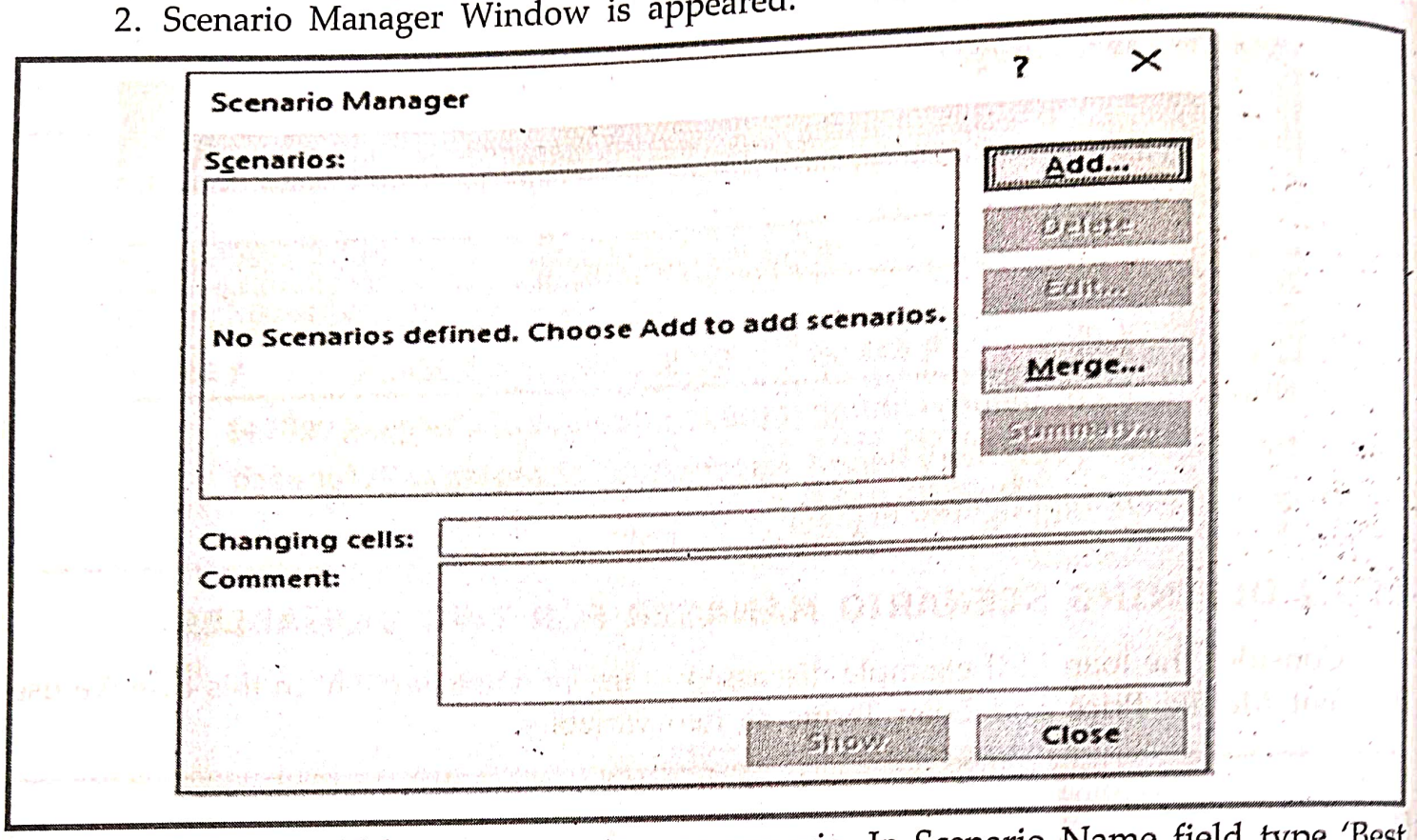
Consider the loan EMI example discussed in the previous section. In this case we use different Mobile Price and Loan Terms as two variables.



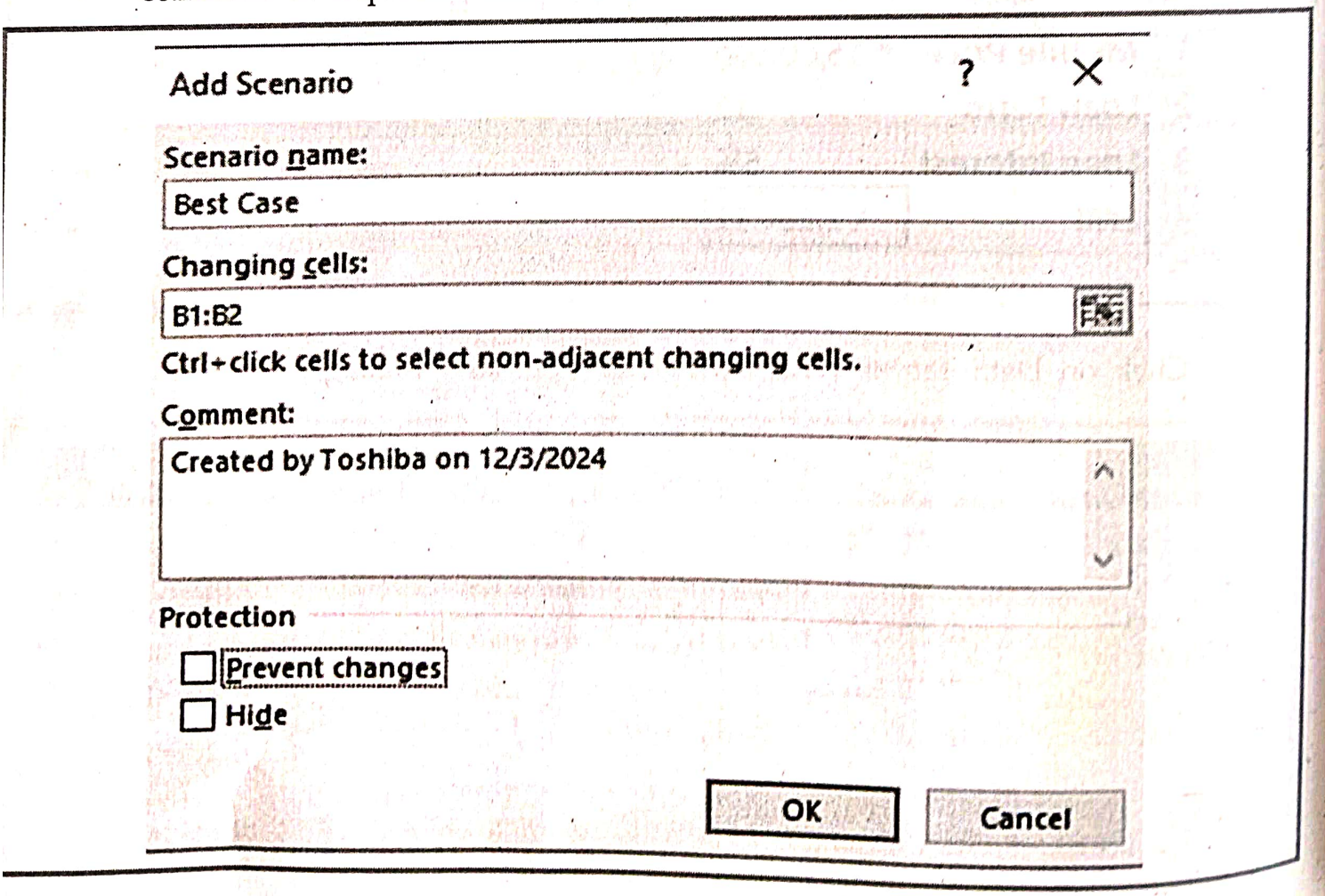
1. Click on Data Tab -> What-if Analysis-> Scenario Manager



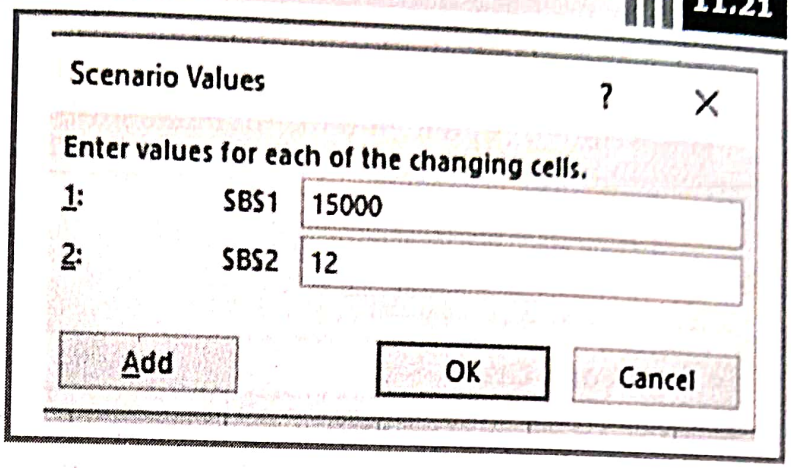
2. Scenario Manager Window is appeared.



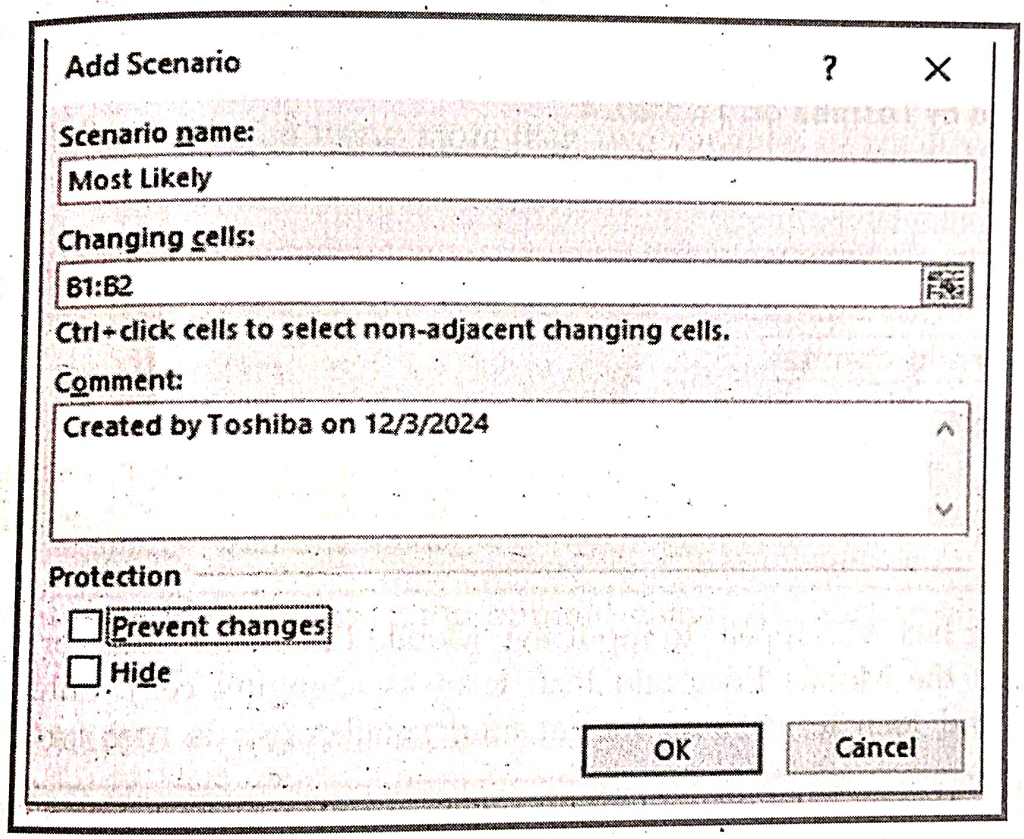
3. Click on Add button to create new scenario. In Scenario Name field type 'Best Case', Changing Cell type 'B1:B2' (Mobile Price and Loan Term) and write comment if required. Click on OK button.



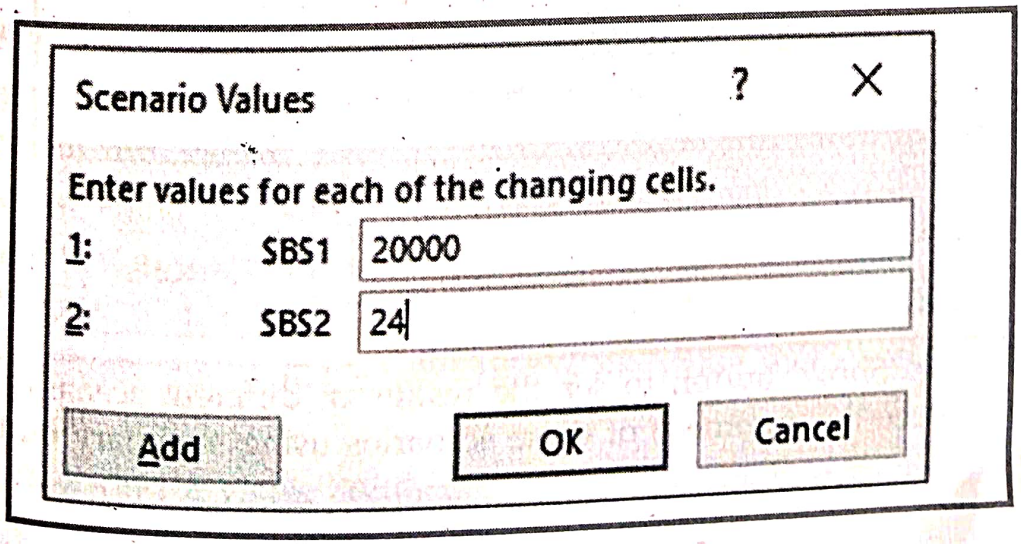
4. A dialog box will appear to input the 'Mobile Price' and 'loan term' (as we have mention the Mobile Price and loan term as changing cell). Enter Mobile Price 15000 and loan term 12 as we consider small terms is best for EMI.



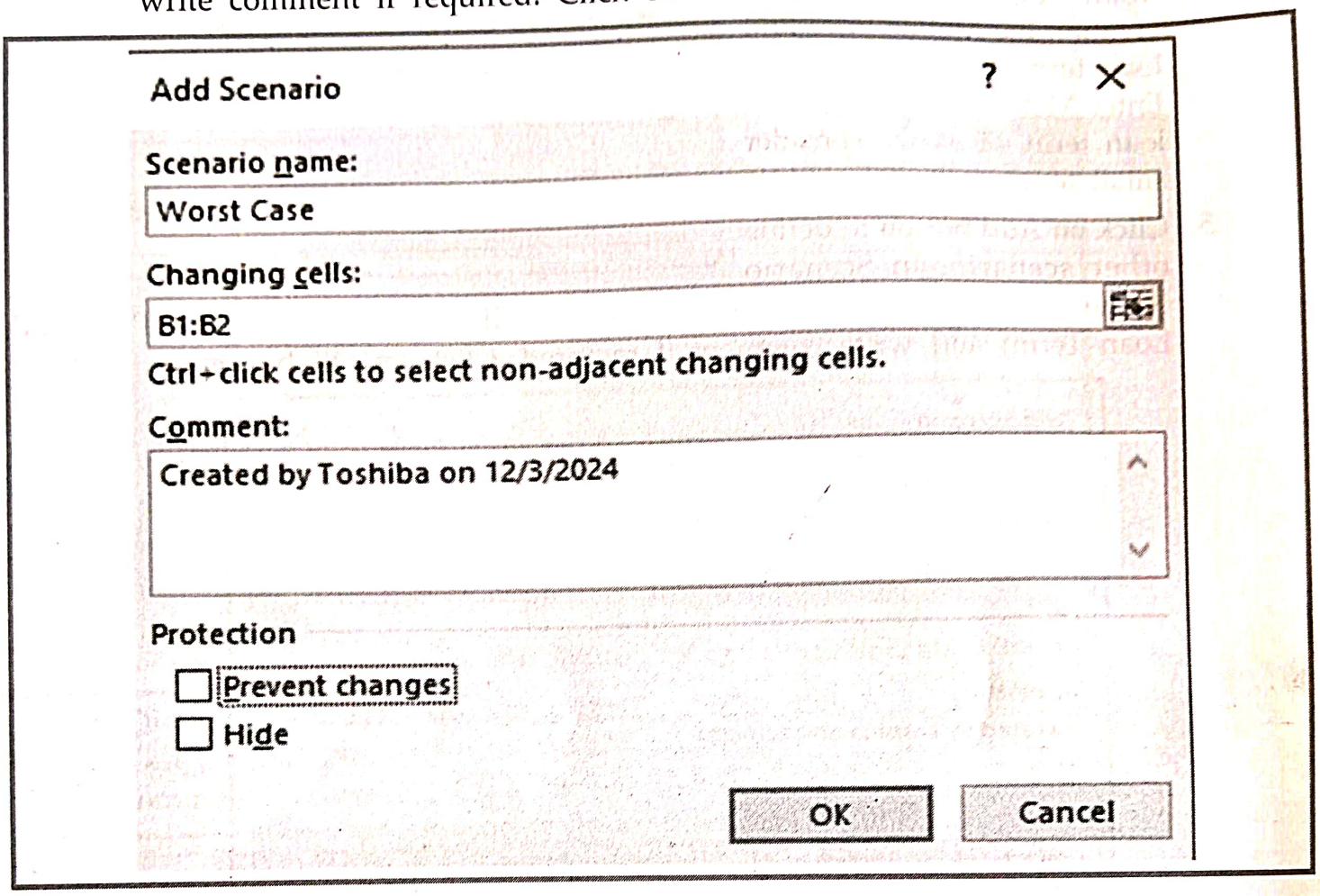
5. Click on Add button to define other scenario. In Scenario Name field type 'Most Likely', Changing Cell type 'B1:B2' (Mobile Price and Loan Term) and write comment if required. Click on OK button.



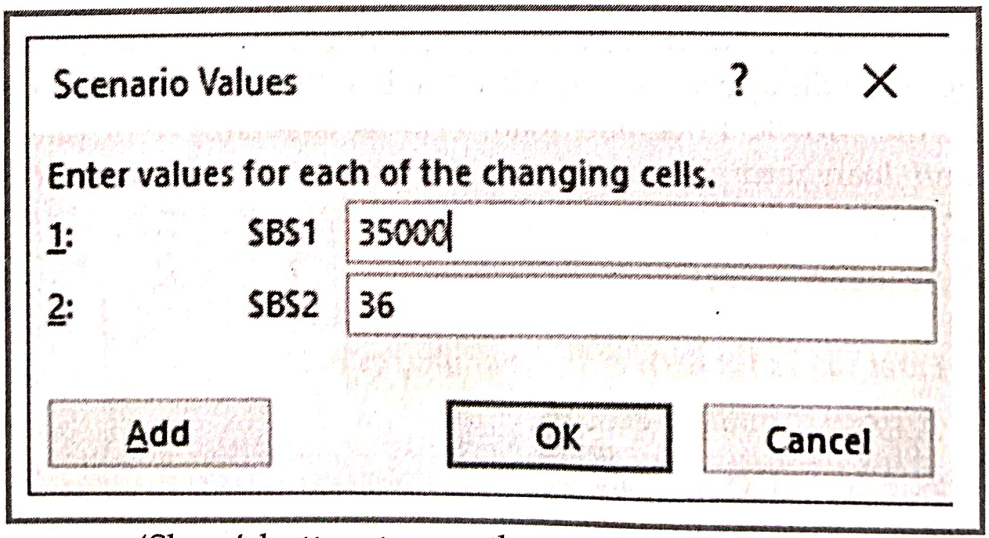
6. A dialog box will appear to input the 'Mobile Price' and 'loan term' (as we have mention the Mobile Price and loan term as changing cell). Enter Mobile Price 20000 and loan term 24 as we consider small terms is best for EMI.



7. Click on Add button to define other scenario. In Scenario Name field type 'Worst Case', Changing Cell type 'B1:B2' (Mobile Price and Loan Term) and write comment if required. Click on OK button.



8. A dialog box will appear to input the 'Mobile Price' and 'loan term' (as we have mention the Mobile Price and loan term as changing cell). Enter Mobile Price 35000 and loan term 36 as we consider small terms is best for EMI.



9. You can use 'Show' button to see the results of different scenarios by selecting it. You can see the summary of these scenarios using summary button as shown below.

Scenario Summary				
	Current Values:	Best Case	Most Likely	Worst Case
Changing Cells:				
B1	₹ 15,000.00	₹ 15,000.00	₹ 20,000.00	₹ 35,000.00
B2	36	12	24	36
Result Cells:				
B4	₹ 449.56	₹ 1,284.11	₹ 877.43	₹ 1,048.98

Notes: Current Values column represents values of changing cells at time Scenario Summary Report was created. Changing cells for each scenario are highlighted in gray.

You can define scenarios using more than two variables or multi variables as per the requirement.

11.4 GOAL SEEK

Scenario manager is useful for known variables, but it may also be used to work backwards. Goal Seek works differently than Scenario Manager by starting with the desired result and calculating the input value to achieve it.Goal Seek is a valuable tool, but it can only handle one independent variable and one dependent variable at a time. To achieve specific aims in a model with multiple independent and dependent variables, discover the optimal combination of values for a subset of the independent variables.

11.4.1 WORKING OF GOAL SEEK

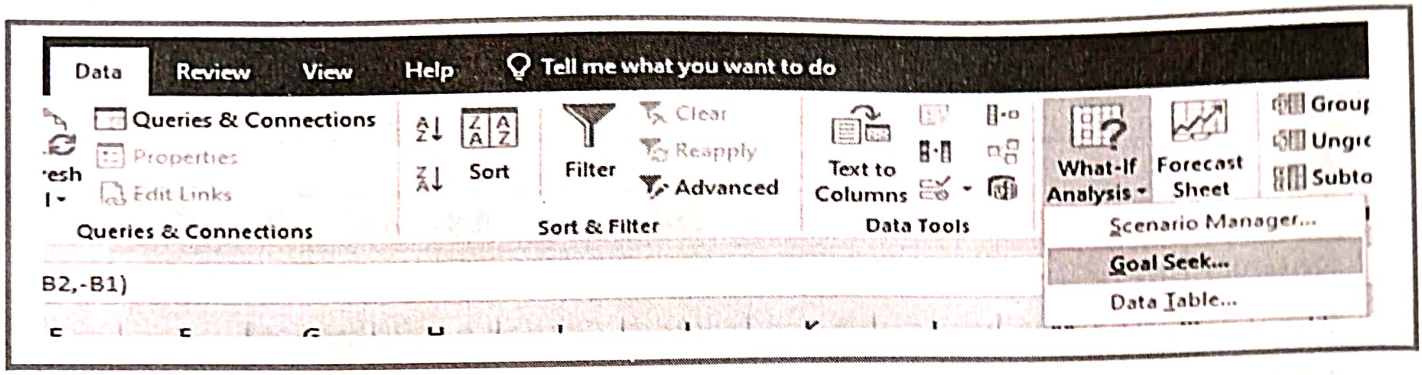
To utilize Goal Seek, in a worksheet, place the formula in one cell and the variable (with an initial value) in another.Goal Seek allows for manipulation of only one variable at a time, even if your formula has numerous variables. Goal Seek employs an iterative approach to identify a solution. Goal Seek tries the variable's initial value to determine if it provides the desired effect. If not, Goal Seek explores different values until it finds a solution.

Excel's goal-seeking function can assist in answering these questions if the situation fulfills the parameters below:

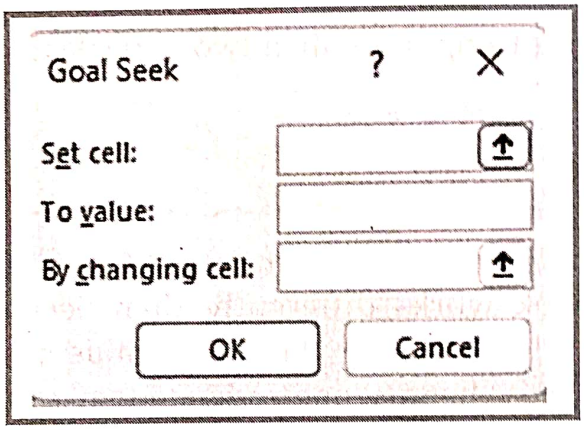
- Only one result cell, which contains a formula. You want to ensure that the result of this computation satisfies the objective you set. Optimization refers to the method by which Excel finds the value that meets this aim.
- There is only one variable cell. This is the value that Excel adjusts to meet your aim in the result cell.
- There's always a viable solution.

11.4.2 USING GOAL SEEK

To use Goal Seek, click on Data Tab -> What-If Analysis-> Goal Seek then Goal Seek dialog box appears.



Following figure shows the Goal Seek dialog box. .



Goal Seek dialog box has following 3 fields:

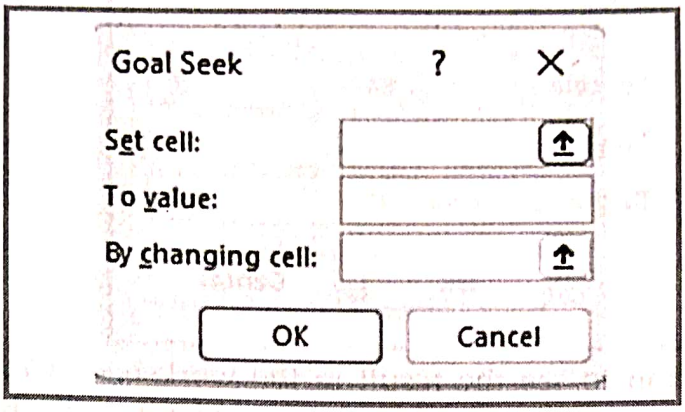
- **Set cell:** Input the cell reference which contain the formula.
- **To value:** Input the target value you want to achieve. Getting this value is purpose of using Goal Seek dialog.
- **By changing cell:** This field has the cell reference for which value will be changed to get the Target value mention in the 'To value'.

Let's take an example of our previous EMI data.

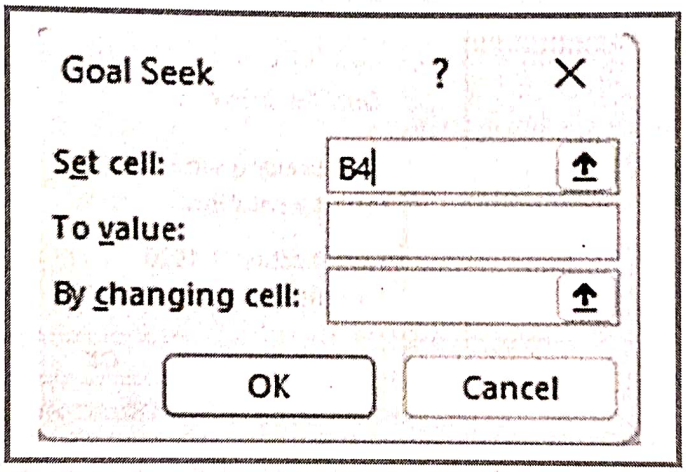
	A	B
1	Mobile Price	₹ 15,000.00
2	Loan Term	12
3	Loan Interest	5%
4	EMI	₹ 1,284.11

Now if customer has constraint on EMI and he want to purchase the mobile accordingly then Goal Seek is appropriate to solve his problem

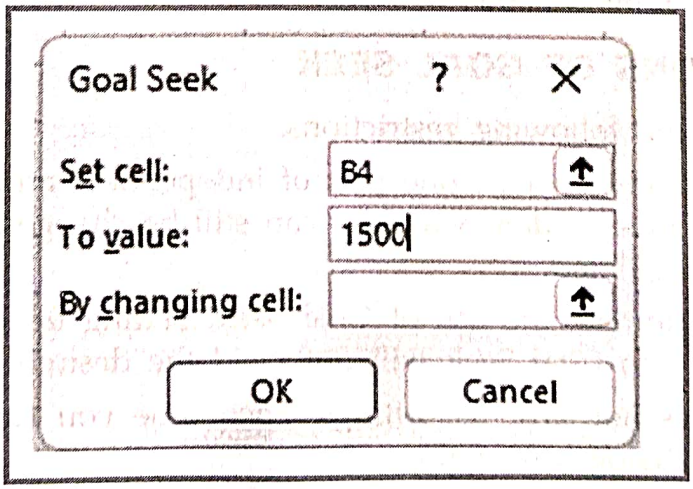
1. Open the Goal Seek dialog box using Data Tab.



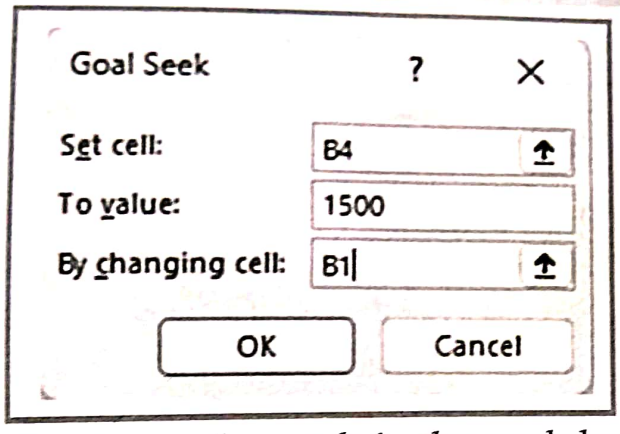
2. 'Set Cell' field requires the cell reference where value will be changed. In this example Type B4 in this field in which we are calculating EMI using PMT formula.



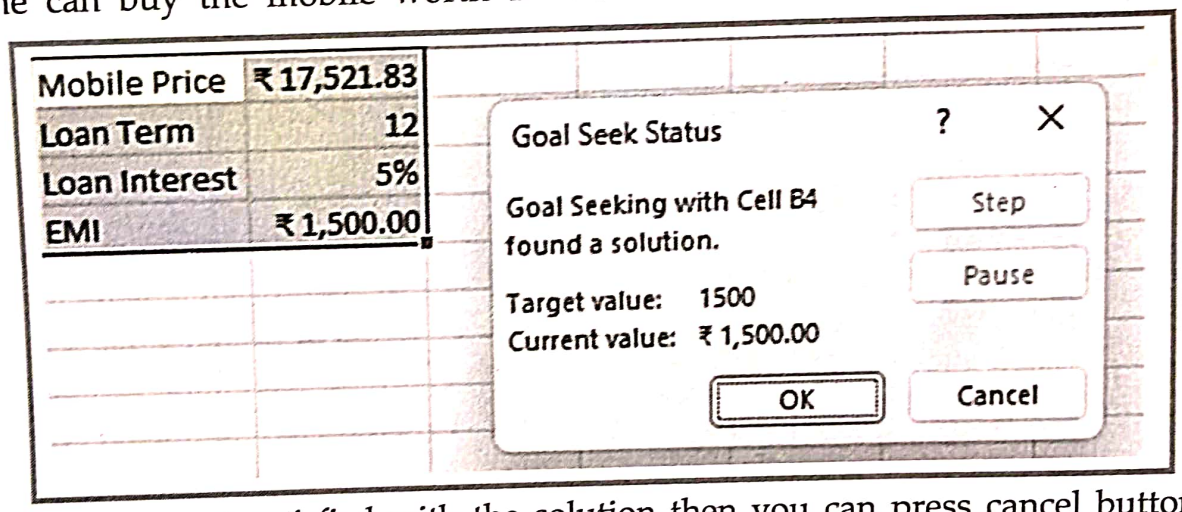
3. Next in 'To value' field input the target value or result you want to achieve. In this example we have to enter the EMI value of choice of customer. Type 1500 as an EMI for this field.



4. Last field is 'By changing cell', it requires cell reference in which value will be updated to achieve the target value. In this example Mobile Price value will be updated then type B1 in this field.



5. Click OK button to see the result in the worksheet. Goal Seek Status will show the status of target value. In this example Mobile price has been changed. It says if customer want to pay EMI of 1500 for 12 months with 5% interest rate then he can buy the mobile worth 17500.



6. If you are not satisfied with the solution then you can press cancel button then excel will revert the changes made in the worksheet and enter new target value if want to do so. If you are satisfied with the solution then press OK button and changes will remain same.

11.4.3 LIMITATIONS OF GOAL SEEK

Goal Seek has some following restrictions.

- The ability to only analyze one pair of independent and dependent variables at a time. Other independent variables can still be changed, but must be manually entered and tabulated.
- You cannot automate the use of Goal Seek. Setting up a model for a user who is unfamiliar with Goal Seek will not yield the desired results.
- Goal Seek does not save its settings. Each time you run it, you must enter all values in the dialog box again.

The Solver resolves some of these issues.

☆☆☆☆☆